Dynamical Filtered Graphs in Finance

Musmeci, Nicolo

Awarding institution:
King's College London

Dynamical Filtered Graphs in Finance

Nicoló Musmeci

Supervisor: Tiziana Di Matteo

Department of Mathematics

King's College London

This dissertation is submitted for the degree of

Doctor of Philosophy

June 2016

To my family

Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university. This dissertation is my own work and contains nothing which is the outcome of work done in collaboration with others, except as specified in the text and Acknowledgements. This dissertation contains fewer than 100,000 words including footnotes.

Nicoló Musmeci

June 2016

Acknowledgements

First of all I want to thank my supervisor Tiziana for her guidance, for encouraging my research and allowing me to grow as a research scientist. I am aware she has always put my interest first and I am really grateful for this. Special thanks to my second supervisor Tomaso as well, who has provided me with invaluable advices, inspiring and encouraging me to explore new ideas and topics.

Many thanks to Raffaello for his constant mentoring help during my first months at KCL: without his support and his valuable friendship it would have been much tougher. I am also very grateful to Paolo for having been such a good friend during these years in London. Likewise, in the second part of my PhD I have really enjoyed the friendship of Riccardo, with whom I have shared ideas, laughs and stress. Among my colleagues I wish also to thank Anshul for being so nice and for the stimulating discussions, not to mention his great help with the proofreading.

Thanks to those friends who I left when I moved to London four years ago, but who have always been there for me: Dario, Flavia, Richard, Pigl, Enzo, Blanco, Matteo, Francesco, Claudia. Our friendship has grown stronger despite the distance and it will last.

I want to thank Adriana for her support and love. In particular I am grateful for her patience during these years of distance, and for thinking it was worth it.

Most of all, I wish to thank my family: my father, whose financial and motivational help has been crucial for finishing (and starting) this PhD; my mother, for her continuous support; and Alice and Vani, who have always been so close to me despite the geographical distance.

Abstract

Financial markets are complex systems characterised by the interaction of several heterogeneous agents. The associated dependence structure is non-trivial and exhibits high levels of non-stationarity and non-linearity. These features make the understanding and forecasting of financial risk very challenging, since regularities observed from historical data do not necessarily mirror future behaviours.

The main aim of this thesis is to investigate the complexity of the dependence structure through network filtering and clustering techniques. We have relied on these tools because they are data driven, model-independent and lend themselves to dynamical analyses. In particular, we have proposed a novel volatility forecasting tool based on network filtering. Furthermore, we have applied the Directed Bubble Hierarchical Tree (DBHT) clustering method for the first time to financial data, highlighting its advantages over other clustering techniques. We have performed statistical hypothesis tests on the dynamical DBHT clustering, in order to track the evolution of each cluster and how their industry-related information is affected by the market regime. We have studied the evolution of correlation-based filtered networks topology by means of data mining and time series techniques, investigating long-term memory properties and their relation with market risk. We have investigated how different measures of dependence perform and compare in terms of network topology, by combining multiplex tools and network filtering for the first time.

We have found that the 2007 financial crisis marks a phase transition between two different regimes of dependence, which display deep dissimilarities in terms of industrial information and remain well distinct for years after the crisis. We have found

that different clustering methods display different sensitivity to these structural changes. Moreover we have shown that correlation-based filtered networks display peculiar patterns in their evolution, notably long-term memory and possibly early-warning signals. After having found that a significant interplay exists between dependence structure variations and volatility, we have introduced a novel volatility forecasting tool which relies on this empirical feature. This new tool overcomes the curse of dimensionality, which limits traditional econometric models to porfolios of few assets. The multiplex analysis has revealed that it is crucial to monitor financial dependence with more than one measure at a time, as linear measures turn out to provide an incomplete picture of the dependence structure, especially during financial crises.

Table of contents

List of figures **15**

List of tables **21**

1 Introduction **23**

2 Financial time series and correlation **31**

2.1 Introduction 31

2.2 Dataset 33

2.3 Financial time series 34

2.3.1 Log-returns 34

2.3.2 Stylized facts of financial time series 36

2.4 Empirical properties of financial correlations 44

2.4.1 Measuring dependence: Pearson coefficient 46

2.4.2 Random matrix theory filtering 49

2.4.3 Subtracting the market mode 53

2.4.4 Dynamical evolution of correlation 54

2.4.5 Economy-related information 58

2.4.6 Time scale 59

2.4.7 The limit of Pearson coefficient: non-linearity in financial correlation 60

2.5 Summary 61

3 Correlation-based filtered networks 63

3.1 Introduction . . . 63

3.2 Financial network: definitions . . . 65

3.2.1 Minimum Spanning Tree . . . 66

3.2.2 Asset Graph . . . 71

3.2.3 Embedded Graphs . . . 75

3.3 Insights from Network-filtering: a brief review . . . 79

3.4 Clustering: a complementary perspective on the dependence structure 82

3.4.1 k-medoids . . . 82

3.4.2 Hierarchical Clustering Methods . . . 83

3.4.3 Linkage methods . . . 85

3.4.4 Directed Bubble Hierarchical Tree . . . 88

3.5 Summary . . . 89

4 Relation between financial market data and real economy 91

4.1 Introduction . . . 91

4.2 Structure and economic information of the correlation clustering . . . 93

4.2.1 Clusters composition . . . 93

4.2.2 Measuring the heterogeneity of clusters size distribution . . . 105

4.2.3 Retrieving economic information: Adjusted Rand Index . . . 108

4.2.4 Retrieving economic information: ICB overexpression . . . 112

4.3 The dynamical evolution of the clustering structure . . . 119

4.3.1 Dynamically retrieving the industrial sectors . . . 123

4.4 Summary . . . 127

5 Evolution of correlation-based networks and clusters tracking 129

5.1 Introduction . . . 129

5.2 Persistence and transitions: dynamical analysis of DBHT . . . 131

5.2.1 A map of structural changes . . . 133

5.2.2 Clusters composition evolution 138
5.3 Memory in the correlation-based network dynamics 144
5.4 Summary 149

6 A new approach to volatility forecasting 151
6.1 Introduction 151
6.2 Data sets: US and UK data 153
6.3 A measure of dependence structure persistence 154
6.4 Dependence analysis 155
6.4.1 Test of significance: block-bootstrapping 157
6.4.2 The advantage of network filtering 158
6.5 Forecasting 160
6.5.1 Measure of forecasting performance 165
6.5.2 Temporal evolution of forecasting performance 172
6.6 Summary 173

7 Multiplex on correlation-based networks 177
7.1 Introduction 177
7.2 Multiplex: a brief introduction 179
7.3 Beyond Pearson coefficient: non-linear measures of dependence . . . 184
7.3.1 Spearman and Kendall correlation 184
7.3.2 Tail-dependence 185
7.3.3 Partial correlation 187
7.4 Data set: non-continuously traded stocks 188
7.5 Multiplex on correlation-based networks 190
7.5.1 A global look at non-linearity: edge overlap 191
7.5.2 A multiplex cartography of network filtering 194
7.5.3 Identifying each contribution: multidegree 196

7.5.4 Interlayer degree-degree correlation: a comparison of assets centrality ranking 200
7.6 Summary 202

8 Conclusions and Outlook **205**

Appendix A DBHT algorithm **211**

Appendix B Bootstrapping **215**

References **217**

List of figures

2.1 **ICB supersectors composition** 33

2.2 **ICB industries composition** 34

2.3 **Prices and returns for Pepsico Inc. (PEP INC.) in the period 1997-2012** 35

2.4 **Tail analysis in the period 1997-2012** 37

2.5 **Tail analysis in the period 1997-2012, grouped by ICB industry**. . 38

2.6 **Aggregation normality for Microsoft stock (MSFT US)** 39

2.7 **Gain/loss asymmetry analysis** 41

2.8 **Prices and log-returns of General Electric stock (GE)** 43

2.9 **Volatility clustering decay exponents**. 44

2.10 **Volatility clustering decay exponents, grouped by ICB industry**. . 45

2.11 **Correlation coefficients distributions**. 48

2.12 **Correlation matrices spectra**. 50

2.13 **Contribution of the assets to each eigenvector**. 52

2.14 **Correlation coefficients distributions for detrended log-returns**. . . 54

2.15 **Correlation matrices spectra for detrended log-returns**. 55

2.16 **Dynamic evolution of average correlation**. 57

3.1 **MST from Pearson correlation among 342 US stocks**. 67

3.2 **Economic information extracted by the MST topology**. 68

3.3 **Degree distribution for the MST**. 69

3.4 **Average neighbour degree as a function of node's degree in the MST**. 70
3.5 **Asset Graph on Pearson correlation among 342 US stocks**. 72
3.6 **Economic information extracted by the AG topology**. 73
3.7 **Degree distribution for the AG**. 74
3.8 **Average neighbour degree as a function of node's degree in the AG**. 75
3.9 **Planar Maximally Filtered Graph on Pearson correlation among 342 US stocks**. 77
3.10 **Economic information extracted by the PMFG topology**. 78
3.11 **Degree distribution for the PMFG**. 79
3.12 **Average neighbour degree as a function of node's degree in the PMFG**. 80
3.13 **Selection of two clusterings from a dendrogram of 7 objects**. . . . 84
3.14 **Dendrogram generated by the Single Linkage method**. 84
3.15 **Dendrogram generated by the Average Linkage method**. 85
3.16 **Dendrogram generated by the Complete Linkage method**. 86
3.17 **Dendrogram generated by the Direct Bubble Hierarchical Tree method**. 87

4.1 **DBHT clustering composition**. 94
4.2 **Single Linkage clustering composition**. 95
4.3 **Average Linkage clustering composition**. 96
4.4 **Complete Linkage clustering composition**. 97
4.5 **k-medoids clustering composition**. 98
4.6 **DBHT clustering composition from detrended log-returns**. 100
4.7 **Single Linkage clustering composition from detrended log-returns**. 101
4.8 **Average Linkage clustering composition from detrended log-returns**. 102
4.9 **Complete Linkage clustering composition from detrended log-returns**. 103
4.10 **k-medoids clustering composition from detrended log-returns**. . . 104

4.11 **Demonstration that different clustering methods show different degrees of disparity in the clustering structure**. 106
4.12 **Demonstration that different clustering methods retrieve to different degrees the ICB industries**. 109
4.13 **Demonstration that different clustering methods retrieve to different degree the ICB supersectors**. 110
4.14 **Amount of ICB information retrieved by the clustering methods, in terms of ICB industries overexpressions**. 113
4.15 **Amount of ICB information retrieved by the clustering methods, in terms of ICB industries overexpressions. Detrended log-returns case**. 114
4.16 **Amount of ICB information retrieved by the clustering methods, in terms of ICB supersectors overexpressions**. 115
4.17 **Amount of ICB information retrieved by the clustering methods, in terms of ICB supersectors overexpressions. Detrended log-returns case**. 116
4.18 **Dynamical evolution of the DBHT clustering**. 120
4.19 **Test of robustness for the dynamical DBHT clustering**. 121
4.20 **Dynamical evolution of the similarity between clustering and ICB**. 124
4.21 **Dynamical evolution of the similarity between clustering and ICB, with detrended log-returns**. 125

5.1 **Dynamical evolution of the DBHT clustering**. 132
5.2 **Persistence analysis based on clustering**. 134
5.3 **Persistence analysis based on metacorrelation**. 135
5.4 **Clusters dynamical composition (part 1)**. 138
5.5 **Clusters dynamical composition (part 2)**. 139
5.6 **Clusters dynamical composition (part 3)**. 140
5.7 **Clusters dynamical composition (part 4)**. 141

5.8 **Analysis of degree evolution for LLTC US Equity**. 146

5.9 **Analysis of clustering coefficient evolution for LLTC US Equity**. . 147

5.10 **Summary of regression analysis for the autocorrelation functions decay**. 148

6.1 **Scheme of time windows setting**. 154

6.2 $ES(T_k, T_{k'})$ **matrices for** $\theta = 1000$**, for NYSE (left) and LSE dataset (right)**. 159

6.3 $\langle ES \rangle(T_k)$ **and** $q(T_k)$ **signals represented for** $\theta = 1000$ **and** $L = 100$. . 160

6.4 $z(T_k, T_{k'})$ **matrices for** $\theta = 1000$**, for NYSE (left) and LSE dataset (right)**. 161

6.5 **Partition of data into training and test set**. 164

6.6 **Partition of data into training and test set**. 165

6.7 **Receiver operating characteristic (ROC) curve for the NYSE dataset**. 170

6.8 **Receiver operating characteristic (ROC) curve for the LSE dataset**. 171

6.9 **Fraction of successful predictions as a function of time**. 172

7.1 **Schematic representation of a two-layers multiplex**. 180

7.2 **Number of stocks that are continuously traded in each time window together with their partition in terms of ICB industries.** . . . 189

7.3 **Mean edge overlap evolution in time**. 192

7.4 **Fraction of edges that exist only on one layer: evolution in time**. . 193

7.5 **Comparison among degree evolution on different layers**. 194

7.6 **Industries evolution in the overlapping degree/partecipation coefficient plane (part 1)**. 197

7.7 **Industries evolution in the overlapping degree/partecipation coefficient plane (part 2)**. 198

7.8 **Normalised multidegree for each ICB industry** I**,** $\kappa_I^{\vec{m}}$**, at different times**. 199

7.9 **Interlayer correlation for each pair of layers, at different time windows**. 201

List of tables

2.1 **Summary table of ρ_{ij} statistics** 48

2.2 **Summary table of ρ_{ij}^{R} statistics**. 53

6.1 **NYSE dataset: correlation between $\langle ES \rangle(T_a)$ and $q(T_a)$**. 162

6.2 **LSE dataset: correlation between $\langle ES \rangle(T_a)$ and $q(T_a)$**. 162

6.3 **NYSE dataset: correlation between $\langle z \rangle(T_a)$ and $q(T_a)$** 163

6.4 **LSE dataset: correlation between $\langle z \rangle(T_a)$ and $q(T_a)$**. 163

6.5 **NYSE dataset: Probability of successful forecasting P^{+}**. 168

6.6 **LSE dataset: Probability of successful forecasting P^{+}**. 169

6.7 **NYSE dataset: Area under the curve (AUC), measured from the ROC curve**. 170

6.8 **LSE dataset: Area under the curve (AUC), measured from the ROC curve**. 171

Chapter 1

Introduction

Over the last few decades the understanding of financial markets has benefited remarkably from the application of complex systems tools [1, 2], contributing to the birth of a new discipline called Econophysics [3–5]. New and traditional problems in Economics and Finance have been tackled through innovative approaches, inspired by the ideas and techniques used in Statistical Physics [3, 6]. Instead of imposing an oversimplified model of the reality for the sake of mathematical tractability, this new generation of models and tools aims to preserve the complexity of the real economic phenomena while discarding the redundancy and noise [7, 5, 8]. This is possible through the application of concepts and techniques from areas like Probability Theory [9, 10], Network Theory [11–13] and Machine Learning [14, 15].

In financial markets, the agents' behaviour is known to be highly non-trivial and difficult to predict, making the asset prices depart from the Gaussian assumption in many respects [3, 16, 17]. Econophysics has contributed to unveil such features [3, 6, 18]. In terms of univariate price series, the main signatures of complexity are fat-tails [17, 19, 20] and multi-scaling [21–29] in log-returns [30], which originate from dynamics such as herding and trading at different time scales [22, 31]. As far as the multi-asset interaction is concerned, the dependence structure turns out to be completely

different from what is expected by both uncorrelated [32, 33] and single-factor normal correlated series [34], displaying a nested, rich and heterogeneous structure [34].

The main motivation of this doctoral thesis is to further investigate the complexity of assets' dependence structure from a dynamical perspective. One of the main challenges in the analysis and modelling of financial data is distinguishing between statistical noise and meaningful dependence structure [32, 33, 35–37]. Another relevant issue is non-stationarity [38–42]: if the future does not stick to the regularities observed in the past, how can we use historical data to describe the future? For what concerns the dependence structure among assets, non-stationarity affects the reliability of risk models and portfolio optimisation tools, which often rely on covariance estimation [43, 38]. Econometrics tools such as multivariate GARCH [44] and stochastic volatility models [45] seek to model the evolution of dependencies, but fail to cope with baskets made of more than few assets [46] due to the large number of parameters that need to be calibrated. Another issue, closely related to non-stationarity, is non-linearity [47–50]; financial assets tend to synchronize in periods of negative returns and desynchronize in periods of positive returns [49, 50]. This behaviour, which linear measures of dependence are not able to capture [51], has been often overlooked in risk management [52].

In this thesis we tackle these problems through the lens of Network Theory, by using so-called "correlation-based filtered networks" [53–61] and hierarchical clustering [62, 63] to analyse and model empirical correlation matrices. Correlation-based filtered networks are tools which map correlation matrices into sparse graphs that retain only a subset of entries using some filtering criterion [53–55]. Following this procedure, they are able to retain the backbone of meaningful interactions and get rid of statistical noise at the same time [64]. Examples of such networks are the Minimum Spanning Tree (MST) [65, 53, 34] and the Planar Maximally Filtered Graph (PMFG) [55, 56]. We will refer to these filtering procedures as "network filtering". Hierarchical clustering is a class of unsupervised learning techniques which seek to build a hierarchy of communities

of similar elements, given a distance matrix for these elements [15, 64]. Network filtering and hierarchical clustering turn out to be closely related and complementary [64, 66]. Their application to financial data has provided important insights into the structure and properties of financial markets. Clustering methods and network filtering have been used to study: asset pricing models [34], the sensitivity of the dependence structure to changes in the sampling frequency of returns [67, 68], the hierarchy of relations among industrial sectors [53, 69–72] and the evolution of financial correlation [73–75], with a particular focus on financial crises [76–78]. Also risk management and portfolio optimisation methods have benefited from these techniques. It has been shown that Markwoitz optimisation can be improved through clustering [79]. Besides, correlation-based filtered networks provide a valuable criterion for asset allocation [80, 81].

By using network filtering and hierarchical clustering, we can translate the statistical problem of non-stationary and non-linear correlations into network analyses, for which a number of techniques from Combinatorics [82] and Machine Learning [15] can be applied. In particular, we apply for the first time to financial data the Directed Bubble Hierarchical Tree (DBHT) technique [66, 83], a clustering method with strict connections with the PMFG. Moreover, we compare different dependence measures by using the multiplex framework [84–86], which allows to quantify dissimilarities among networks on the same set of nodes [87–92], increasing the amount of information which can be analysed through Network Theory. Although multiplex networks have been already used in Finance [93–95], this is the first application to the analysis of financial dependence.

The temporal perspective is one of the main focus of this thesis. In particular, we track dynamical DBHT clusters by using statistical tests based on hypergeometric distribution [96], which have been proposed for characterising communities in heterogeneous complex systems [97, 98]. We rely on similarity measures between clustering data, such as the Adjusted Rand Index [99], to distinguish between different market regimes

[39]. We use regression techniques [15] to investigate the influence of the average market dynamics (market mode) [68] on the evolution of correlation. We analyse the evolution of correlation-based filtered networks topology through time series [30] and data mining [15] techniques. Notably, we use metrics from Network Theory to build a predictive model which forecasts market volatility of portfolios made of hundred of assets, overcoming the limitations of traditional econometric methods [46].

The thesis is organised as follows. In Chapter 2, we discuss the most relevant properties of financial time series and of their dependence structure. To this end, we perform statistical analyses on an equity data set of 342 US stocks daily prices. We first focus on the univariate properties of financial series, introducing those signatures of complexity known as "stylized facts" of financial series [6, 17, 100, 101]. Then we turn to the multivariate structure of dependence [102]. Our empirical analyses highlight the mixture of meaningful signal and noise present in the empirical correlation matrix [32, 33, 35]. In particular, we analyse the correlation from a dynamical perspective by means of the exponentially smoothed Pearson estimator [103], highlighting how financial crises deeply affect the dependence structure. Interestingly, we find that both market-level and inter-sector correlations played a role in these structural changes in the 2007 financial crisis.

In Chapter 3 we introduce network filtering [53, 55, 54] as a powerful tool for investigating further these features. We summarise the main types of correlation-based networks used in Finance, in particular the MST [65, 53, 34] and the PMFG [55, 56]. We review the main insights that network filtering has provided over the last 15 years into Finance [3]. Moreover, we elaborate on the connection between network filtering and hierarchical clustering [63], introducing Linkage methods [62] and the DBHT [66]. We apply both network filtering and hierarchical clustering to the equity dataset, demonstrating that they extract topological structures which are significant in terms of underlying economic activity. Also we show how their topologies display a hierarchical

and heterogeneous organisation that is typical of complex networks [13], as revealed by the analyses on different network metrics [12, 13].

We rely on these powerful filtering tools in Chapter 4, which is devoted to investigate the relation between dependence structure and industrial sector activity, through the information provided by the hierarchical clustering. We quantify such relation by comparing the clustering based on correlation with the industrial partition, through metrics such as the Adjusted Rand Index [99] and the hypergeometric hypothesis test [96, 97]. Performing this comparison dynamically and after subtracting the average market returns, we show how such a relation is affected by turbulent market periods and the market mode. Additionally, we consider and compare different clustering methods. All these analyses are of interest for those investment strategies which rely on clustering techniques on correlation matrices [79, 104, 105]. Compared to previous analyses on filtering of financial correlation [106], this study has the advantage of not relying on any assumptions about the returns distribution.

We investigate further the dynamical evolution of dependencies in Chapter 5, which deals with the non-stationarity of financial correlation from a network filtering perspective. We introduce a new measure of similarity between dependence structures at different periods, based on the Adjusted Rand Index [99] among DBHT clusterings at different time windows. We use this measure to quantify and study the rate of change of the dependence structure. Our approach is intuitive and model independent, unlike traditional stationarity tests [38]. The results we obtain indicate that diversification strategies based on industrial membership have become less effective after the 2007 financial crisis. We also report for the first time evidence of long range memory in the evolution of correlation-based filtered networks, reflecting a phenomenon analogous to the volatility clustering in log-returns [100].

These results open interesting scenarios for the forecasting of financial correlation. In Chapter 6 we take the first step in this direction by introducing a novel tool for predicting market volatility variations. Specifically, we introduce a new measure,

the "dependence structure persistence", which quantifies the rate of change of the dependence structure and it can be easily computed from dynamical correlation-based networks. We show that such a measure provides information on future volatility variations, and propose to take advantage of this relation by using a classification technique [14] to predict whether next year volatility will increase or decrease. The power of this novel tool is proved through an out-of-sample analysis [15] on two different datasets. These analyses are the first step towards the application of network filtering to modeling and forecasting, beyond the descriptive analyses it has been mostly used for so far. Moreover, they make forecasting the volatility of entire markets possible, representing a remarkable improvement over traditional volatility forecasting models [46].

Finally, in Chapter 7 we tackle the problem of non-linearity in the dependence structure [49, 50] through the multiplex framework [84, 85]. The idea is to investigate the dissimilarity between networks computed from both linear and non-linear dependence measures [107, 108]. These differences are well captured by the multiplex frameworks and provide insights into the extent of non-linearity in the dependence structure. We find that financial crises widen the differences among the measures of dependence, just when evaluation of risk becomes of the highest importance. Overall, our results highlight the importance of monitoring financial risk by means of more than one measure of dependence at the same time. To the best of our knowledge, our analyses represent the first application of multiplex to correlation-based filtered networks, as well as the first investigation into the non-linearity of dependencies from both a global and dynamical perspective.

I list here the peer reviewed articles I have produced in the course of my PhD. The articles are listed in chronological order of appearance:

1. N. Musmeci, T. Aste and Di Matteo, T., "Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Method", PLoS ONE 10 (4), 2015, e0126998. doi: 10.1371/journal.pone.0126998.

2. N. Musmeci, T. Aste and Di Matteo, T., “Risk diversification: a study of persistence with a filtered correlation-network approach”, Journal of Network Theory in Finance 1, 2015, 1-22.

3. N. Musmeci, T. Aste and Di Matteo, T., “What does past correlation structure tell us about the future? An answer from network filtering”, Scientific Reports, submitted, 2016.

4. N. Musmeci, V. Nicosia, T. Aste, Di Matteo, T. and V. Latora, “The multiplex structure of financial markets”, Scientific Reports, submitted, 2016.

5. R. J. Buonocore, N. Musmeci, T. Aste, and T. Di Matteo. “Two different flavours of complexity in financial data”, EPJ ST, submitted, 2016.

In the course of my PhD I have presented my research in the following conferences and events:

- “Workshop on Econophysics and networks across scales”, Lorentz Center, Leiden, Netherlands, 27-31st May 2013.
- “London Graduate School PhD Day 2013”, LSE, London, UK, 1st March 2013.
- “5th Workshop on Complex Networks CompleNet”, Bologna, Italy, 12-14th March 2014.
- “Open Statistical Physics”, Open University, Milton Keynes, UK, 26th March 2014.
- “GENED Workshop: Networks in Finance and Macroeconomics”, Institute for the World Economy, Kiel, Germany, 28-29 April 2014.
- “Statistical Mechanics of Glassy and Complex Systems”, King’s College London, London, UK, 19-20th May 2014.
- “London Graduate School PhD Day 2015”, LSE, London, UK, 13th March 2015.

- “ComplexiLIMS seminars”, London Institute for Mathematical Sciences, London, UK, 27th May 2015.
- “Econophysics Colloquium 2015”, Prague, Czech Republic, 14-18th September 2015.

Chapter 2

Financial time series and correlation

In this chapter we review the most relevant properties of financial time series and investigate their dependence structure. In particular, we discuss fat-tails, autocorrelation, cross-correlation and non-stationarity. We also introduce a data set of equity data on which we perform a set of preliminary empirical analyses. Part of the results presented in this chapter has also been published in the paper "Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Methods" in 2015 [109].

2.1 Introduction

Time series analysis is an important part of the study of many complex systems. From ecosystems to financial markets, from social networks to the climate, time series (such as temperatures, prices, number of organisms etc.) are indeed the main quantitative output that we can extract from these systems. Within this broad set of analysis, the study of dependencies among different time series is especially valuable, as it gives insight into the mutual interactions of different parts of the complex system.

The dependence structure among assets returns is of interest in Finance for several reasons [110]. First of all, every risk assessment - both of a small portfolio and the whole market - needs to take into account how different assets prices move together; in

general, higher correlations are associated to higher risks, and measures of dependence are heavily used in Portfolio Optimization and Risk Management [111–113]. Besides, dependencies have become much relevant for trading too, as proven by the number of derivatives whose price depends on correlation [114–117].

Several studies have shown that dependencies in financial markets display characteristic features, mirror of the underlying complexity of the market agents interactions [118, 113, 119–121]. Tools such as correlation-based networks and random matrix theory have revealed that the dependence structure is characterized by both high level of noise and a backbone of meaningful information [53, 76, 55, 57–60, 70, 32, 33, 35–37, 76]. Such structure contains a certain degree of information related to the real economy and it is highly non-stationary [38]; its evolution synchronizes with the overall trend of the market, in particular during financial crises [119]. Many of these empirical facts are at odds with some assumptions underlying traditional econometric tools (such as Capital Asset Pricing Model [122]) and are the foundation for a new generation of models.

In this chapter we review the main empirical properties of financial time series and dependence structure. To illustrate these properties we analyse a data set of 342 US daily stock prices over a period of 15 years. In particular we show that financial time series in the equity data set depart from the assumption of uncorrelated normal random variables; notably they display complex temporal and cross-sectional dependence structure, as well as a certain degree of non-stationarity.

This chapter is organized as follows. In Section 2.2 we describe the data set and its main characteristics; in Section 2.3 we review the main univariate properties of financial time series and we show that the data set reproduces such properties; in Section 2.4 we focus on the dependence structure of financial assets, we review the existing literature on financial correlation and we analyse the dependence structure of the equity data set.

2.2 Dataset

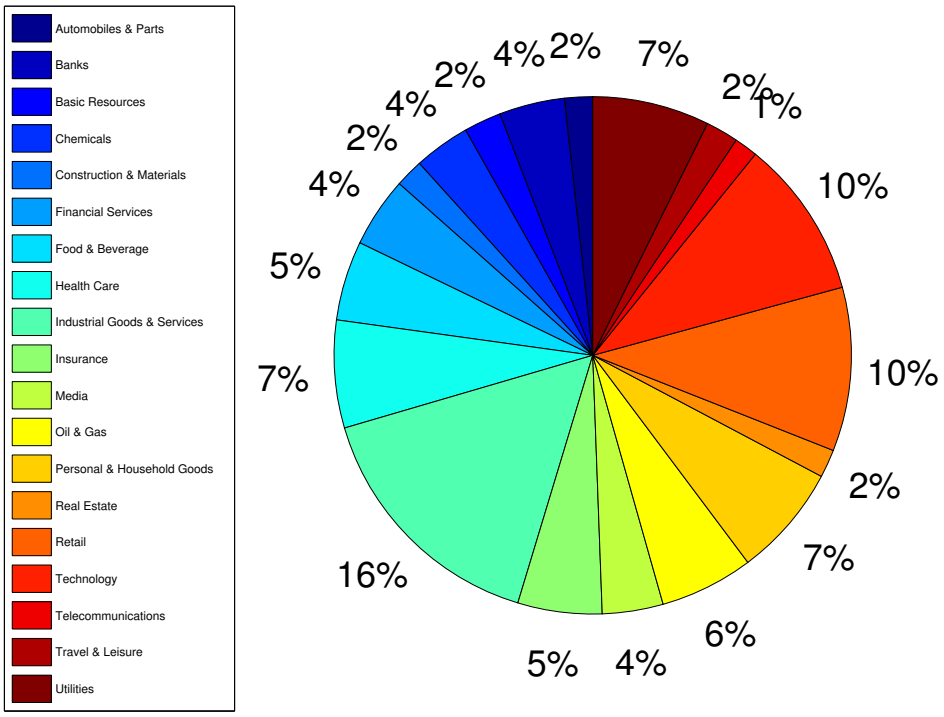

Fig. 2.1 **ICB supersectors composition** Pie chart showing the composition of the entire set of stocks in terms of ICB supersectors.

The original analyses we will present in this chapter are performed on a dataset of equity data provided by Bloomberg. It is composed by daily closing prices of 342 US stocks, covering 15 years from 02/01/1997 to 31/12/2012. This period covers a number of significant events that have characterised the market evolution, notably the Dot-com bubble [123] and the 2007-2008 financial crisis [124].

All stocks have been continuously traded throughout this period of time. The set of stocks has been chosen in order to provide a significant sample of the different industrial sectors in the market. We have chosen the ICB industrial classification, that yields 19 different Supersectors, that in turns gather in 10 Industries: the percentages of stocks belonging to each ICB supersector and industry are reported in Figs. 2.1 and 2.2.

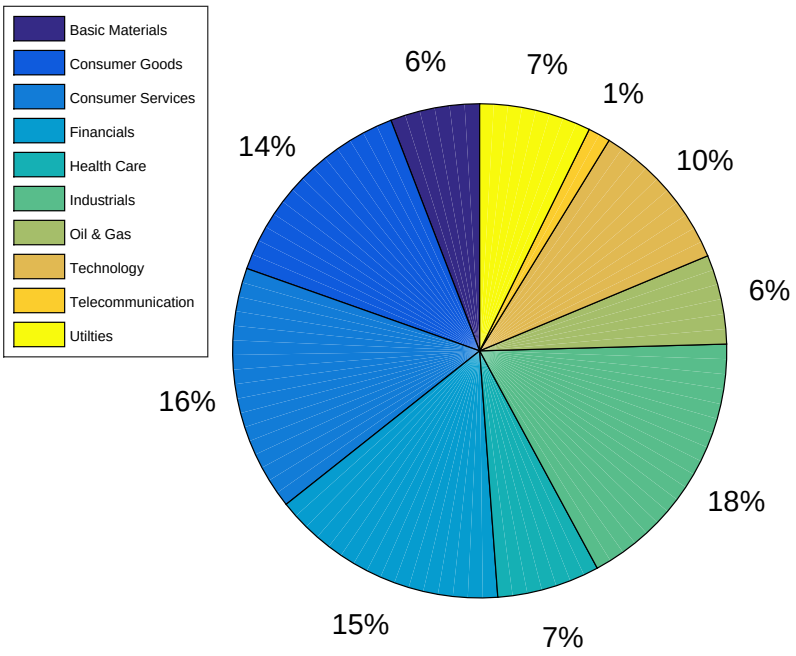

Fig. 2.2 **ICB industries composition**. Pie chart showing the composition of the entire set of stocks in terms of ICB industries.

2.3 Financial time series

In this section we review the main empirical properties of financial time series. We also perform a set of statistical analyses on the equity data set to confirm the validity of such properties. We will focus on the univariate features; the issue of dependence between different assets will be treated in the Section 2.4.

2.3.1 Log-returns

Defining the variables of interest is the first step in any scientific analysis. Let us denote with $P_i(t)$ the price of an asset i at time t. It turns out that $P_i(t)$ is not the best choice for statistical analyses and modeling, due to its non-stationarity and long-range autocorrelation [30]. This is confirmed by applying the Kwiatkowski–Phillips–Schmidt–Shin (KPSS) test for stationarity [125] to the data set: we have found that the test is rejected for all 342 stocks prices with a p-value lower than 0.01.

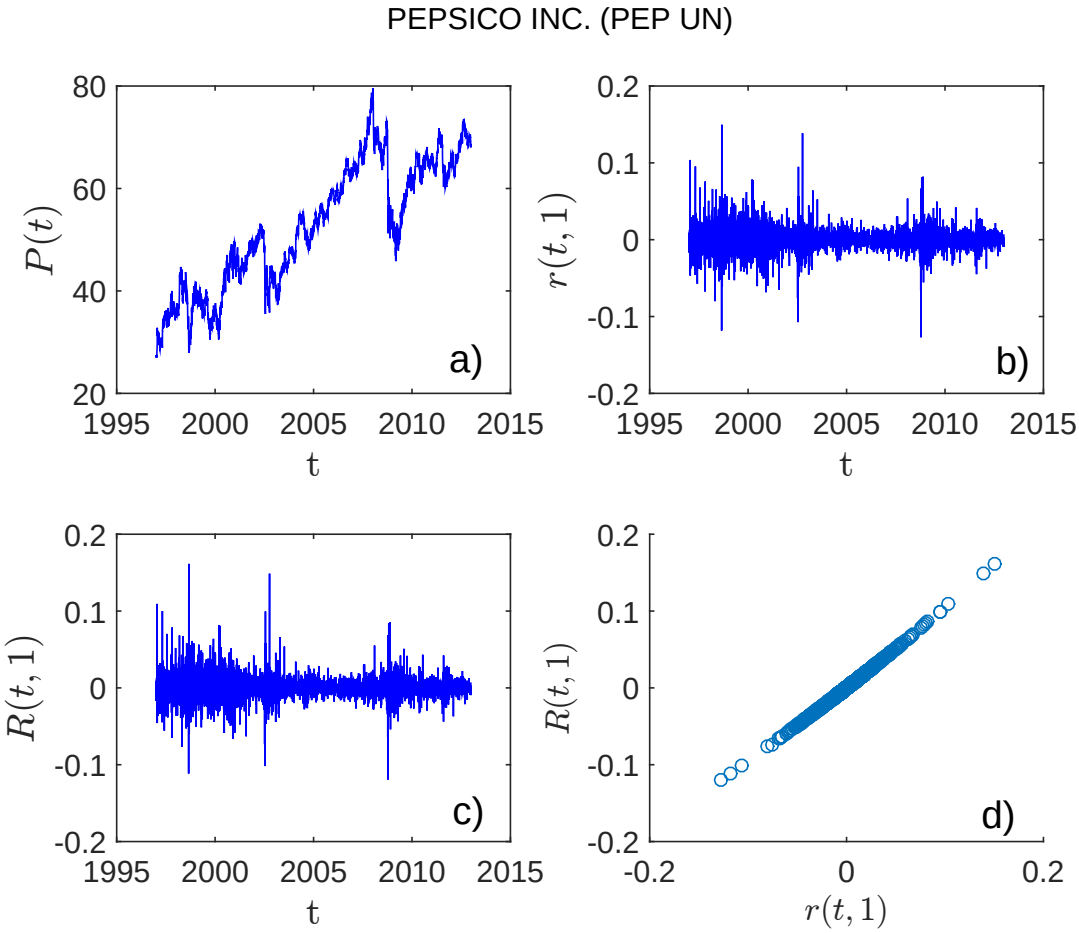

Fig. 2.3 **Prices and returns for Pepsico Inc. (PEP INC.) in the period 1997-2012**. a) Price $P_i(t)$; b) log-returns $r_i(t,\tau)$, with $\tau = 1$ day; c) simple relative increment $R_i(t,\tau)$, with $\tau = 1$ day; d) Scatter plot of $R_i(t,\tau)$ against $r_i(t,\tau)$.

In order to remove - or at least reduce - such non-stationarity it is convenient to define the log-return at scale τ as [30, 3]:

$$r(t,\tau) = \log(P(t+\tau)) - \log(P(t)) \quad . \tag{2.1}$$

An alternative definition is the simple relative increment at scale τ [30]:

$$R(t,\tau) = \frac{P(t+\tau) - P(t)}{P(t)} \quad . \tag{2.2}$$

By means of algebra manipulation it is possible to show that there is a simple relationship between $r(t,\tau)$ and $R(t,\tau)$, namely $r(t,\tau) = \log(1+R(t,\tau))$ [30]. In this

thesis we will use log-returns (or simply returns) $r(t,\tau)$: their advantage over $R(t,\tau)$ is their higher statistical tractability, as well as their additive property [30]. However, the two quantities are approximately equal when $R(t,\tau)$ is small, as $\log(1+x) = x + O(x^2)$. In this thesis we will choose $\tau = 1$ day, that corresponds to $R(t,\tau)$ of the order of few percents: in this range log-returns and simple relative increments differ of the order 10^{-3}, that is negligible for the purpose of statistical analysis. As an example we have computed $r(t,\tau)$ and $R(t,\tau)$ for daily prices of stock PEPSICO INC. from the data set (see Fig. 2.3). In Fig. 2.3 d) we show the scatter plot of the two quantities: the similarity is almost perfect, with a correlation of 0.9997.

Since we will focus on daily returns only ($\tau = 1$ day), in the rest of this thesis we will drop τ and indicate log-returns simply with $r(t)$.

2.3.2 Stylized facts of financial time series

We here review the most important properties of financial time series from a univariate perspective. Since the seminal work by Louis Bachelier in the early 1900 [126], the time evolution of stock prices has been object of study in Economics and Statistics. In [126] Bachelier suggested to use the Brownian motion to model such evolution. Later this assumption turned out to be way too simplistic and we are now aware of some empirical features - known as stylized facts - that make financial time series highly complex and not easy to model. At the present, no model is able to reproduce all these features together, although single features can be replicated by specific models. We here discuss the most relevant of these empirical facts [3, 127].

- **Heavy tails**: the departure from normality is particularly important in the tails of returns distribution. Let us call f_r the unconditional distribution of returns; it turns out that f_r displays fat-tails (or heavy-tails) consistent with a power-law behavior and not compatible with a normal distribution [17, 19]. This power-law scaling is empirically measured from the corresponding cumulative distribution function (cdf):

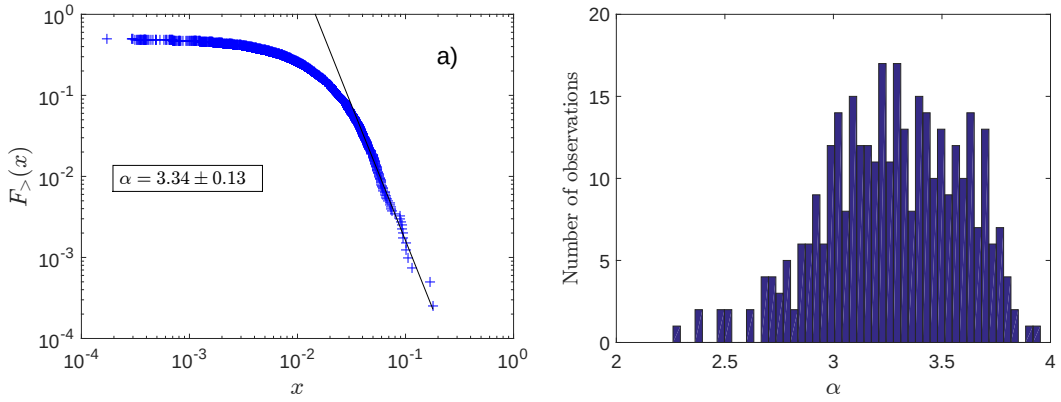

Fig. 2.4 **Tail analysis in the period 1997-2012**. a) $F_>(x)$ in log-log scale for log-returns of Microsoft stock (MSFT US) in the period 1997-2012; the estimated linear fit and correspondent α are shown. b) Histogram of power-law exponents α for all stocks in the data set. All exponents fall between 2.26 and 3.95, with an average of 3.27.

$$F_<(x) = P(r < x) = \int_{-\infty}^{x} f_r dr \ , \tag{2.3}$$

whose complementary function $F_>(x) = 1 - F_<(x)$ shows the following behavior:

$$F_>(x) \sim x^{-\alpha} \ . \tag{2.4}$$

In a distribution with power-law tails events that are many standard deviations away from the mean are more likely than in a normal distribution [8]. The probability of such extreme events depends on the exponent α: the higher α, the closer the distribution is to a normal one. Estimated exponents α for financial assets range between 2 and 5 [19, 20].

We have computed the α exponent for all the 342 stocks in the data set. To this end we have used the so-called rank-frequency plot [128]. Given a set of T observations $\{x_1, x_2, ..., x_T\}$, this method allows to estimate the cdf by simply computing their ranking normalised by T: indeed we have $Rank(x_i)/T = 1 - F_<(x_i)$ [128]. The exponent α and its error are then estimated through a linear fit in log-log scale of $F_>(x_i)$ against x_i. The lowest extremum x_{min} at which the fit

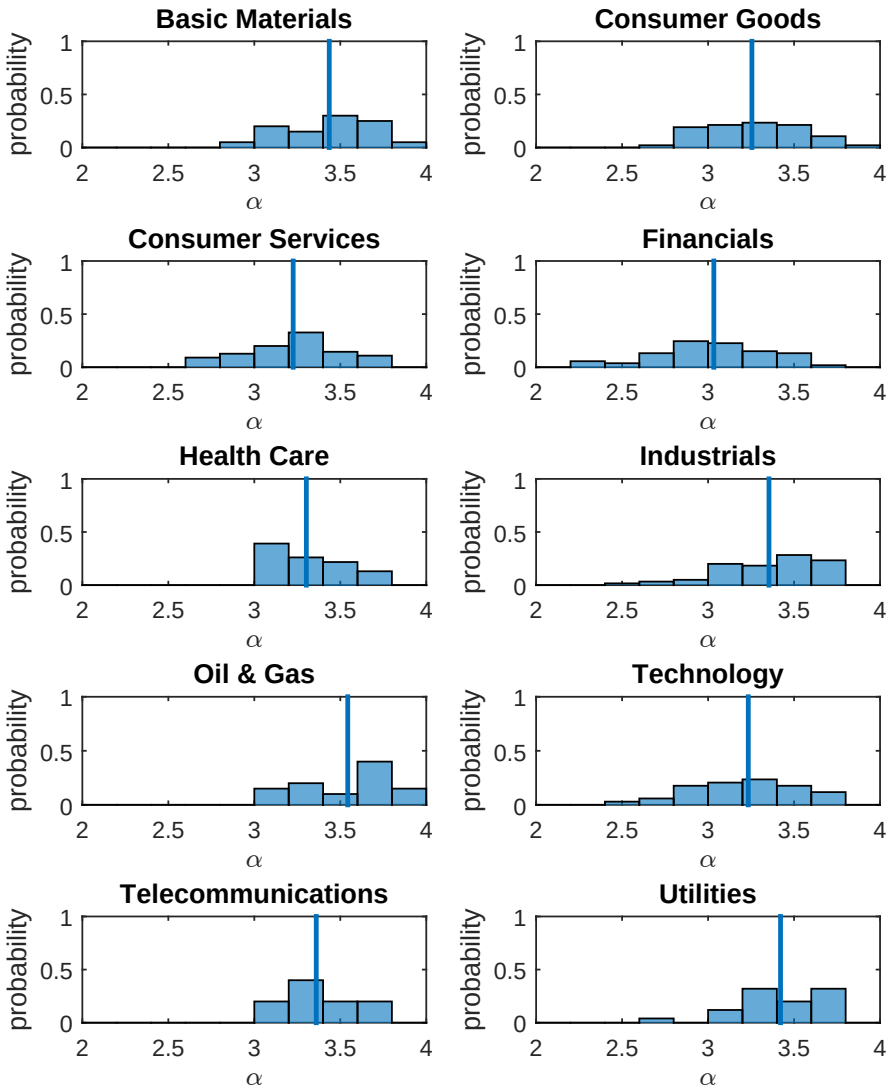

Fig. 2.5 **Tail analysis in the period 1997-2012, grouped by ICB industry.** Histogram of power-law exponents α, grouped by ICB industry. The industry with largest mean α is Oil & Gas and Utilies have the largest mean α, whereas Finance displays the strongest departure from normality with the lowest mean α.

is started is chosen by performing a different fit for each x_{min}: the chosen x_{min} is the value that provides the best fit in terms of Kolmogorov-Smirnov goodness-of-fit statistic [128]. In Fig. 2.4 a) log-log plot of $F_>(x)$ is shown for Microsoft stock in the period 01/1997-12/2012 as an example, with correspondent α estimation. In Fig. 2.4 b) the α values obtained from the entire data set of 342 stocks are

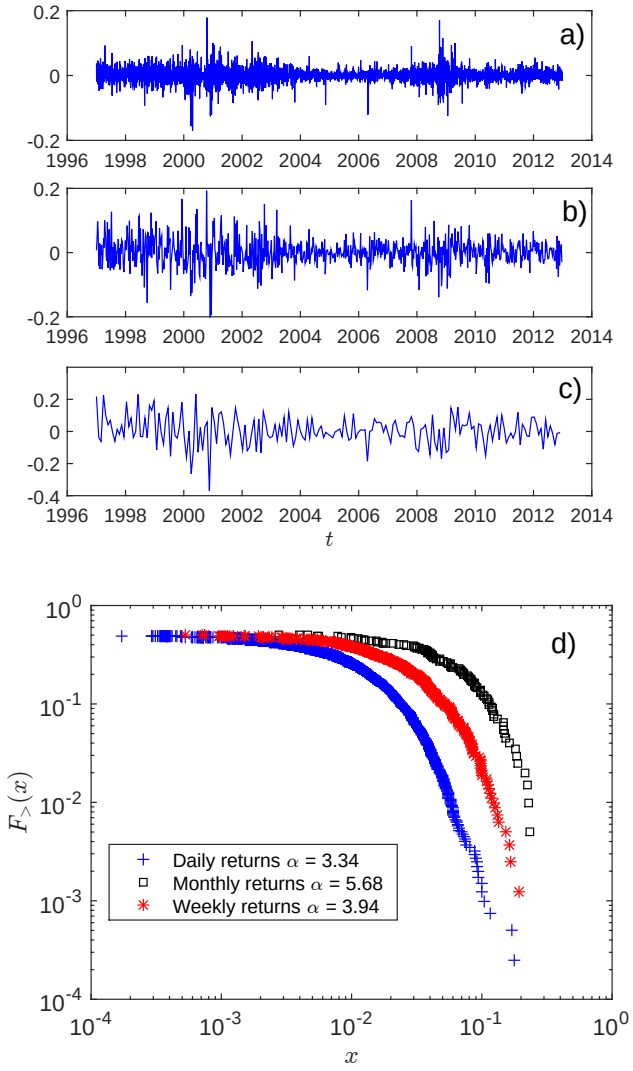

Fig. 2.6 **Aggregation normality for Microsoft stock (MSFT US)**. Log-returns for MSFT US at daily a), weekly b) and monthly c) frequency, over the period 1997-2012. d) $F_>(x)$ in log-log scale for log-returns for daily, weekly and monthly returns of MST US, with correspondent α estimations. It is visible the gradual convergence from power-law to exponential (normal-like) tails, quantified by the increase in α.

shown in an histogram: all values fall between 2.26 and 3.95, with a mean of 3.27. A certain variability among ICB industries can be observed as well, as shown in Fig. 2.5 where we plot normalised histograms for ICB industry. The industry with largest mean α is Oil & Gas, with $\bar{\alpha} = 3.54$, followed by Utilities with $\bar{\alpha} = 3.42$; whereas Finance displays the strongest departure from normality, with $\bar{\alpha} = 3.03$.

Standard models for option pricing, such as Black and Scholes model, do not take into account such fat-tails feature of asset returns: this discrepancy between theory and empirical facts originates the so called volatility smile in option pricing [129].

- **Aggregational normality**: when the time horizon τ increases, the unconditional distribution approaches a normal distribution, as a consequence of the Central Limit Theorem [130]. The theorem applies as long as $\alpha > 2$ (as it is the case for all stocks in the analysed data set), since this implies that the log-returns variance is defined. We have checked this effect by calculating log-returns at three different time scales (daily, weekly and monthly) for Microsoft stock in the period 01/1997-12/2012, and then estimating the corresponding α. As shown in Fig. 2.6, α increases with τ, making the distribution closer and closer to a normal distribution.

- **Gain/loss asymmetry**: returns distribution is typically slightly left-skewed; this means that large losses tend to be more likely than large gains [101]. In the equity data set we have estimated the skewness for each stock by means of the estimator s, that given a set of T observations $\{x_1, x_2, ..., x_T\}$ reads [131]:

$$s = \frac{\frac{1}{T}\sum_{i=1}^{T}(x_i - \bar{x})^3}{(\frac{1}{T-1}\sum_{i=1}^{T}(x_i - \bar{x})^2)^{\frac{3}{2}}} \quad . \tag{2.5}$$

  In Fig. 2.7 we show with an histogram the results. About 68% of stocks display negative sample skewness, with lowest value equal to $s_{min} = -13.6$. On the contrary, there are no large positive values, with all positive s falling below 1.1.

- **Absence of autocorrelation**: daily returns show no significant autocorrelation, even at lag 1. The autocorrelation at lag l is defined as follows:

$$\gamma_i(l) = \frac{\langle r_i(t+l)r_i(t)\rangle - \langle r_i(t+l)\rangle\langle r_i(t)\rangle}{\sqrt{\langle r_i(t)^2\rangle - \langle r_i(t)\rangle^2}\sqrt{\langle r_i(t+l)^2\rangle - \langle r_i(t+l)\rangle^2}} \quad , \tag{2.6}$$

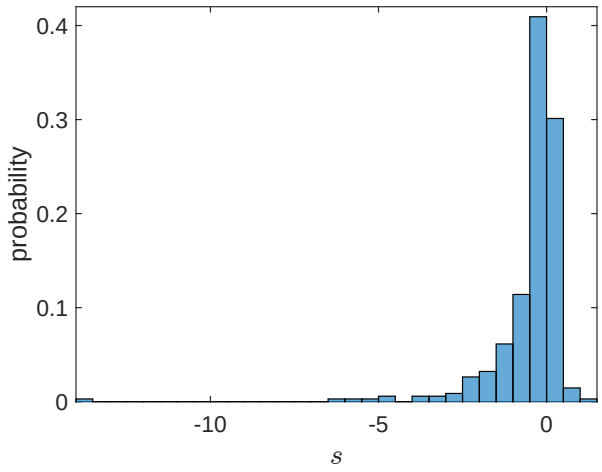

Fig. 2.7 **Gain/loss asymmetry analysis**. Histogram of empirical skewness s for all stocks in the data set. 68% of stocks have negative value of skewness, that means negative returns are more likely than positive returns.

where $\langle ... \rangle$ indicates the expected value. The uncorrelation at all lags is consistent with the so called Efficient-Market Hypothesis [132]: correlated returns would indeed represent arbitrage opportunities for traders. We have estimated the autocorrelation by using the corresponding sample estimator [30]:

$$\hat{\gamma}_i(l) = \frac{\frac{1}{T-l}\sum_{t=1}^{T-l}(r_i(t)-\bar{r}_i)(r_i(t+l)-\bar{r}_i)}{\sigma_i} \quad , \tag{2.7}$$

where T is the total number of observations, $\bar{r}_i = 1/T\sum_t r_i(t)$ and $\sigma_i = \sqrt{1/(T-1)\sum_t (r_i(t)-\bar{r}_i)^2}$. We show as an example the sample autocorrelation function $\hat{\gamma}_i(l)$ for General Electric stock in the period 01/1997-12/2012 in Fig. 2.8 c). Already at $l = 1$ the autocorrelation is consistent with absence of autocorrelation within the confidence interval (blue horizontal lines).

Slight negative autocorrelation is observed only at small intra-day scales, due to microstructure effects [3, 127].

- **Volatility clustering**: uncorrelation does not imply independence, and in fact returns at different times turn out to be uncorrelated but dependent [3]. In

particular non-linear functions of returns, such as absolute values and squares, are highly autocorrelated and the decay shape of such autocorrelation is roughly power-law [17, 100]. This phenomenon is called volatility clustering since it appears as clusters of highly volatile days and is one of the main manifestation of long-term memory in financial time series. It partially accounts for the heavy tails in the returns distribution; however we observe heavy tails even after taking into account this effect (e.g. through GARCH models) [127].

From the equity data set we have computed for each stock the sample autocorrelation function of the absolute value of returns:

$$\hat{\gamma}_i^{abs}(l) = \frac{\frac{1}{T-l}\sum_{t=1}^{T-l}(|r_i(t)| - \overline{|r_i|})(|r_i(t+l)| - \overline{|r_i|})}{\sigma_i^{abs}} \quad , \tag{2.8}$$

where $\overline{|r_i|} = 1/T\sum_t |r_i(t)|$ and $\sigma_i^{abs} = \sqrt{1/(T-1)\sum_t (|r_i(t)| - \overline{|r_i|})^2}$. In Fig. 2.8 d) we show $\hat{\gamma}_i^{abs}(l)$ for General Electric stock in the period 01/1997-12/2012. One can observe a significant autocorrelation even at $l = 250$ days, that is one year. In Fig. 2.8 b) the log-returns for the same stock are shown, from which we can observe how periods of high/low fluctuations tend to group together.

In order to quantify the rate of decay of $\hat{\gamma}_i^{abs}(l)$ with l we have estimated the β exponent in the power-law relation:

$$\hat{\gamma}_i^{abs}(l) = l^{-\beta} \quad . \tag{2.9}$$

To this end we have performed a linear fit in log-log scale of $\hat{\gamma}_i^{abs}(l)$ against l for each stock in the data set. We summarise the results in Fig. 2.9, where we show the histogram of β. The values range between 0.22 and 0.63, with a mean of 0.41. In Fig. 2.10 we show histograms of β for each different ICB industry; as for the fat-tail exponent, we find heterogeneity among different industries. In

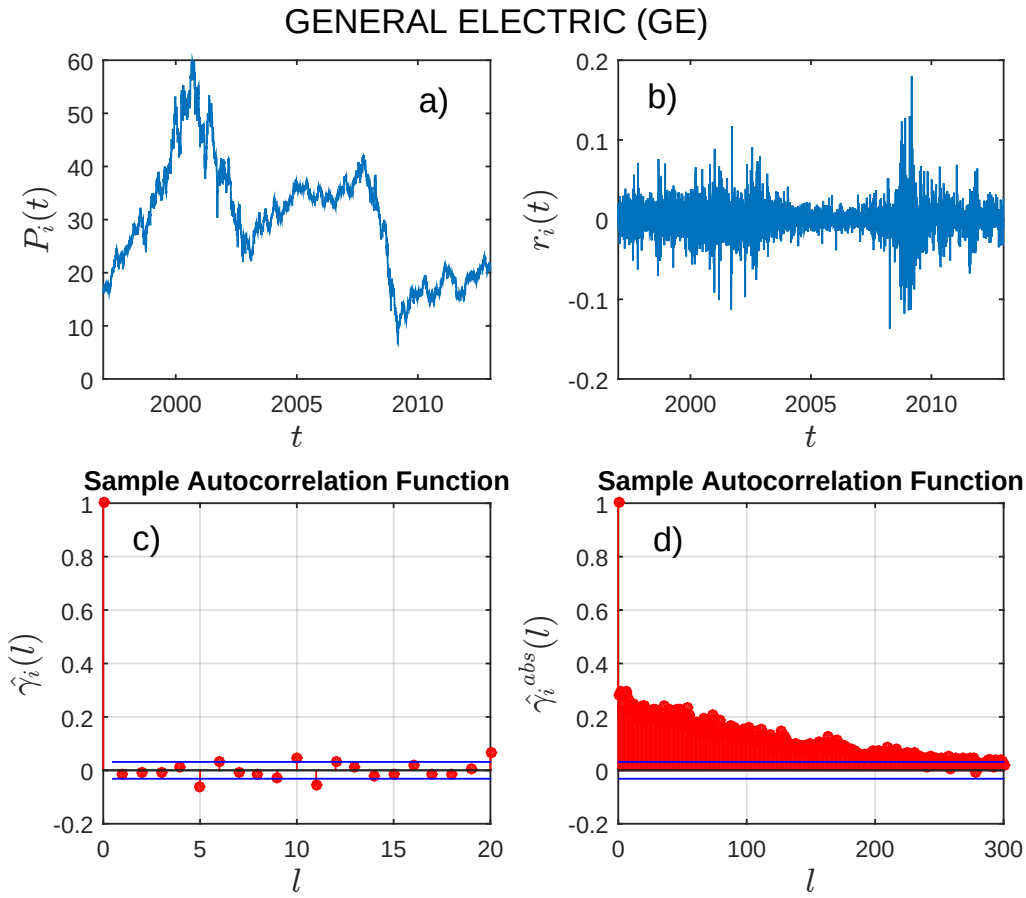

Fig. 2.8 **Prices and log-returns of General Electric stock (GE)**. a) Prices $P(t)$ in time; b) log-returns r_t in time: clusters of high and low fluctuations are visible; c) sample autocorrelation function of log-returns, where it is evident the lack of significant autocorrelation already at lag 1; d) sample autocorrelation function of absolute value of log-returns: significant autocorrelation is evident still at $lag > 200$, due to the cluster volatility structure observed in b).

particular we find that stocks in Technology display on average the slowest decay in autocorrelation, having the lowest average β (0.325).

- **Leverage effect**: volatility is negatively correlated with returns; that is, periods of negative returns show often high volatility. The most prevalent explanation is the following: companies whose stock prices go down tend to become automatically more leveraged (as their equity part is decreasing while the debt is constant), therefore they become riskier and stock prices more volatile [133, 134]. Several

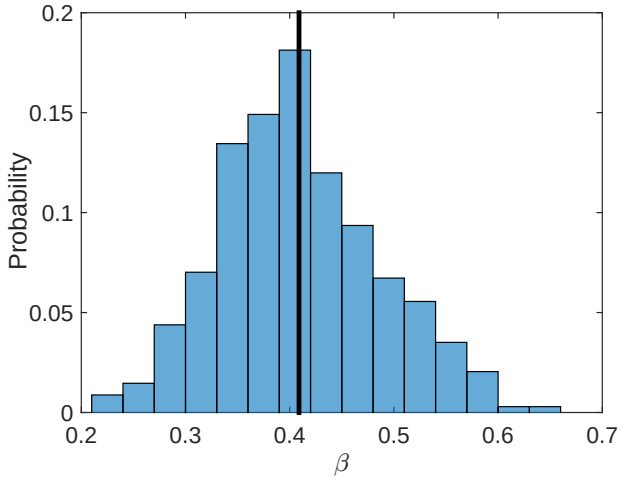

Fig. 2.9 **Volatility clustering decay exponents**. Histogram of decay exponent β for all stocks in the data set. The values range between 0.22 and 0.63, with a mean of 0.41.

models have been proposed to describe the leverage effect, mainly based on stochastic volatility models [135].

- **Volume/volatility correlation**: trade volume is positively correlated with the volatility [136].

To summarise, in this section we have reviewed the main statistical properties of financial log-returns. To this end we have performed a set of analyses on the data set of equity daily prices. We have shown how equity log-returns display a level of complexity, in terms of extreme events and memory, that distinguishes them from the Brownian motion assumption. In the next section we will discuss how complexity arises in the interaction among different assets by focusing on the analysis of the dependence structure.

2.4 Empirical properties of financial correlations

As a matter of fact, log-returns of different assets display a high cross-dependence, even across industries and asset classes [119]. After all, market participants typically

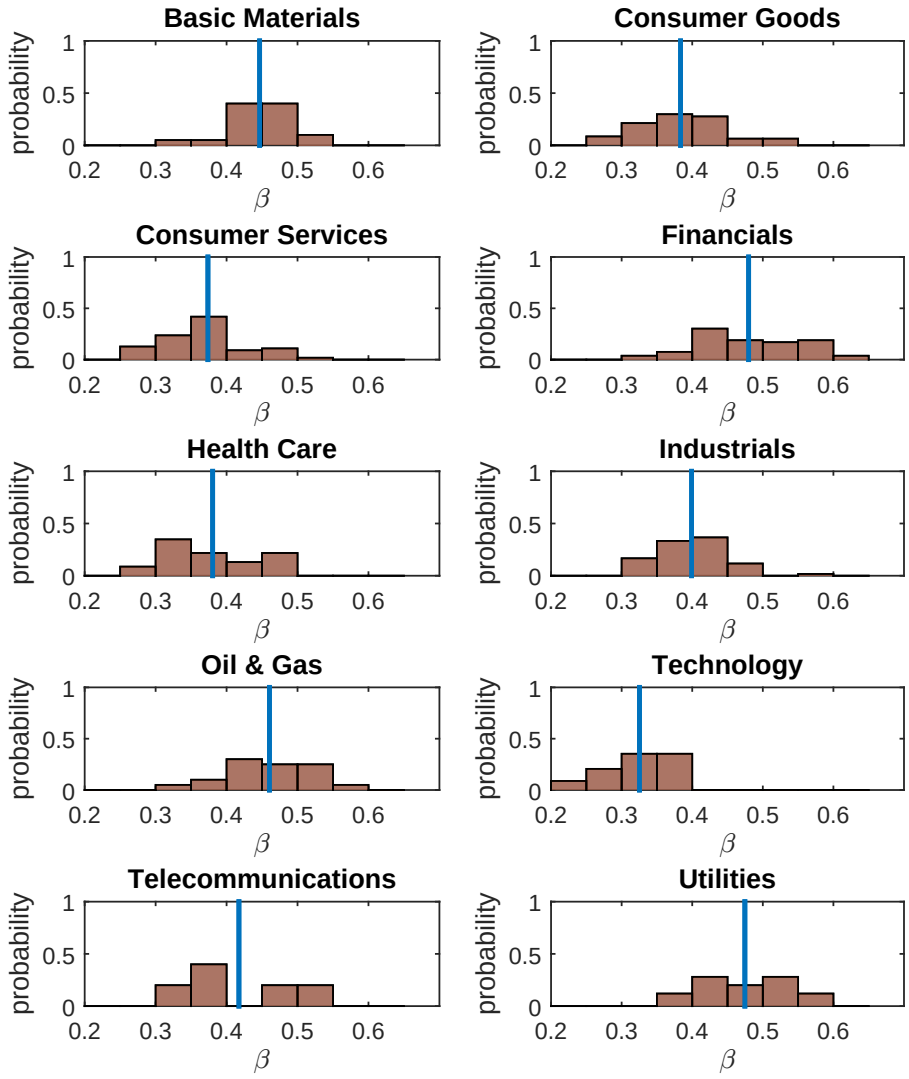

Fig. 2.10 **Volatility clustering decay exponents, grouped by ICB industry**. Histogram of decay exponent β for all stocks in the data set, grouped by ICB industry. Stocks in Technology display on average the slowest decay in autocorrelation, having the lowest average β.

trade and invest in more than one asset, making price movements synchronize or anti-synchronize. The resulting structure of dependencies is of great interest in Finance. Assessing accurately such structure is a key issue in Risk Management as well as in Pricing and Trading. As a consequence, temporal evolution of this structure is crucial too, and with the expression "Correlation risk" we refer to risk arising from potential changes in correlation [110].

In this section we review the main empirical findings concerning financial dependencies. To this end we will also analyse the dependence structure of the data set. We begin by discussing how to quantitatively measure the dependence between two random variables.

2.4.1 Measuring dependence: Pearson coefficient

The measure of dependence between two random variables is one of the most widespread problems in Probability and Statistics. Measures of dependence are used in virtually every field where a rigorous analysis of data is required, from Biology and Physics to Finance and Sociology. After all, the first quantitative studies on the topic appeared in an applicative context, as they were first introduced in Biometrics by Francis Galton [137] and Karl Pearson [102] at the end of the Nineteenth century.

Pearson coefficient was the first index of dependence to be introduced [102] and is still one of the most popular in the applications. Given two assets i and j with log-returns $r_i(t)$ and $r_j(t)$, the Pearson coefficient is defined as follows:

$$\rho_{ij} = \frac{Cov(r_i, r_j)}{\sigma_i \sigma_j} \quad , \tag{2.10}$$

where $\sigma_i = \sqrt{Var(r_i)}$, $\sigma_j = \sqrt{Var(r_j)}$ and $Cov(r_i, r_j) = E[(r_i - \langle r_i \rangle)(r_j - \langle r_j \rangle)]$ is the covariance between the two series of returns. The covariance is a multivariate generalization of the variance, and provides a measure of to what extent r_i and r_j vary together; however it is not a suitable dependence measure since it is not normalized and therefore not easily interpretable [138]. Pearson correlation ρ_{ij} can therefore be seen as a normalised covariance, that ranges between $+1$ (perfect correlation) and -1 (perfect anti-correlation) [138]. It can be estimated from a sample of observations $\{r_i(t)\}$ and $\{r_j(t)\}$, with $t = 1, ..., T$, by using the Pearson estimator [102]:

$$\widehat{\rho}_{ij} = \frac{\sum_t (r_i(t) - \overline{r}_i)(r_j(t) - \overline{r}_j)}{\sqrt{\sum_t (r_i(t) - \overline{r}_i)^2} \sqrt{\sum_t (r_j(t) - \overline{r}_j)^2}} \quad . \tag{2.11}$$

An intuitive interpretation of ρ_{ij} is provided by its connection with the theory of linear regression. Indeed it turns out that ρ_{ij}^2 can be re-written as follows [51]:

$$\rho_{ij}^2 = \frac{\sigma_j^2 - min_{a,b}E[(r_j - (ar_i + b))^2]}{\sigma_j^2} \quad , \tag{2.12}$$

where $min_{a,b}E[(r_j - (ar_i + b))^2]$ can be interpreted as the residual variance that is left after the best linear fit is performed on r_j (using r_i as independent variable). Since the parameters a and b are chosen in order to minimise the square of residuals, "best" is meant in terms of Ordinary Least Squares. ρ_{ij}^2 is therefore the fraction of variance that is explained by the linear regression [51]: if all the variance of r_j can be explained by the linear model, than $\rho_{ij}^2 = 1$. On the other hand, if $\sigma_j^2 = E[(r_j - (ar_i + b))^2]$ for each combination of parameters a and b, then a linear model is not able to explain even part of the variability of r_j, and $\rho_{ij}^2 = 0$. In theory of regression the square of the Pearson estimator $\hat{\rho}^2$ is called "Coefficient of determination" and is indeed used as an index of goodness of fit [139]. In this sense Pearson coefficient can be seen as a test of linear dependence between r_j and r_i [138].

Let us compute the Pearson estimator in Eq. 2.11 on the equity data set. To this end we have taken the entire set of $T = 4025$ observations, covering the whole period 1997-2012. Since we have $N = 342$ stocks, we obtain $N(N-1)/2 = 58311$ distinct Pearson coefficients. A summary of the main statistical properties of this population is shown in the first row of Tab. 2.1 (off-diagonal entries only). As we can see, the mean is more than three standard deviations greater than zero and the maximum value is close to 1. Moreover a positive skewness of 0.584 indicates a fatter tail in the region of positive coefficients. A remarkably different sample is obtained by randomly shuffling the series of log-returns in Eq. 2.11: by doing so we destroy any meaningful comovement present in the original, aligned set of returns [140]. A statistical summary of a correlation matrix obtained in this way is shown in the second row of Tab. 2.1: the average collapses to a value very close to zero, the standard deviation decreases, the asymmetry almost disappears and the distribution shrinks around zero. A visual

Table 2.1 **Summary table of ρ_{ij} statistics**. We report the main sample features for the Pearson coefficient computed on log-returns (first row) and on randomly shuffled log-returns (second row).

Type	Mean	Standard Deviation	Skewness	Min	Max
log-returns	0.303	0.097	0.584	- 0.027	0.9706
shuffled log-returns	9.2556e-05	0.0158	0.005	-0.0760	0.0931

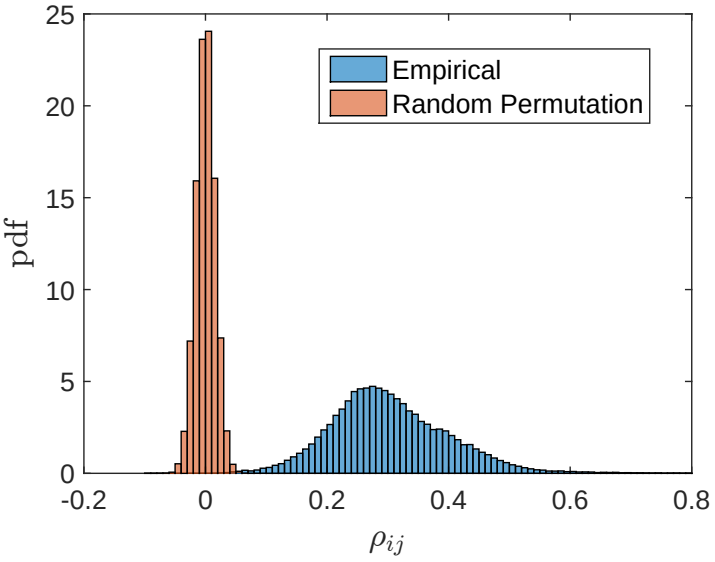

Fig. 2.11 **Correlation coefficients distributions**. Upper graph: histogram of correlation matrix entries $\rho_{ij}(T)$ for empirical (blue) and shuffled log-returns (orange). The empirical distribution has its mean shifted towards positive values and has higher standard deviation. Upper graph: same comparison for correlation matrix entries $\rho^R_{ij}(T)$ detrended of the market mode. Although the empirical mean is now back to zero, empirical distribution is more spread and skewed towards positive values.

comparison of the two samples is in Fig. 2.11. These differences point out the presence of significant dependence structure in the empirical correlation matrix, structure that disappears only when comovements among stocks are removed.

Pearson correlation is very popular for a number of reasons. It can be easily calculated and is formally elegant [51]. Moreover it is easy to manipulate under linear transformations, and allows to calculate the variance of any linear combination of random variables, making it a valuable tool for portfolio theory [113]. Thirdly, it can be

shown that Pearson coefficient is the natural measure of dependence for multivariate normal distributions and, more in general, elliptic distributions [51].

It allows quite simple inference and confidence intervals computation as well. Indeed, under the assumption of bivariate normal distribution for r_i and r_j it is possible to show that the measured correlation coefficient $\hat{\rho}$ in Eq. 2.11 is distributed according to the following distribution [138]:

$$P(\widehat{\rho},\rho,T) = \frac{1}{\pi}(T-2)(1-\widehat{\rho}^2)^{\frac{T-4}{2}}(1-\rho^2)^{\frac{T-1}{2}}\int_0^{+\infty}\frac{dr}{(\cosh r-\widehat{\rho}\rho)^{T-1}} \quad , \qquad (2.13)$$

where T is the length of $\{r_i(t)\}$ and $\{r_j(t)\}$ samples and ρ is the true correlation coefficient given by Eq. 2.10. Through this distribution we are able to test e.g. the null hypothesis $\rho = 0$ in a sample, or to estimate confidence intervals for ρ. We have run this test for the Pearson coefficients computed in the data set, and it turns out that only 345 out of 58311 pairs of assets - about 0.006 % - fail to reject the null hypothesis $\rho = 0$; this result is especially remarkable since the significance level α for each test has been chosen equal to 0.01 with the conservative Bonferroni correction for multiple tests [141–143], namely $\alpha = 1/(N(N-1)0.5)$.

2.4.2 Random matrix theory filtering

Hypothesis tests on correlation indicate that the great majority of Pearson coefficients in the data set are significantly different from zero. However, this fact does not imply that the correlation matrix is unaffected by statistical noise. To recognize it we have to change perspective and analyse the spectrum of the correlation matrix. Introduced in the context of financial correlation in 1999 [144, 145, 32, 33, 36, 35, 73], the Random Matrix Theory (RMT) provides an analytical expression for the asymptotic distribution $P(\lambda)$ of the eigenvalues (spectrum) of a set of uncorrelated, standardized normal random

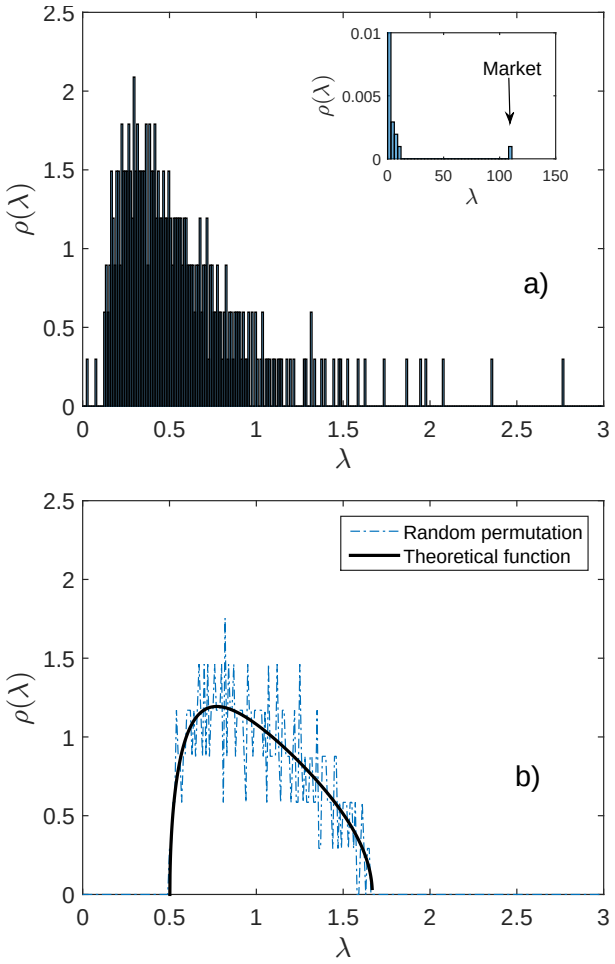

Fig. 2.12 **Correlation matrices spectra**. a) Eigenvalue distribution of correlation matrix $\rho_{ij}(T)$ from log-returns. The inset plot shows the same distribution at a larger scale, to include the largest eigenvalue. b) Eigenvalue distribution of correlation matrix from shuffled log-returns, with theoretical distribution expected from uncorrelated series (black solid curve). As we can see the empirical distribution is in good agreement with the model, that implies shuffling has destroyed the dependence structure of $\rho_{ij}(T)$.

variables. Specifically, calling N the number of variables and T their length, in the limit $N \to \infty$ and $T \to \infty$ with $Q = T/N > 1$ and fixed, we have [145]:

$$P(\lambda) = \frac{Q}{2\pi} \frac{\sqrt{(\lambda_+ - \lambda)(\lambda - \lambda_-)}}{\lambda} \quad , \tag{2.14}$$

where

$$\lambda_\pm = 1 + \frac{1}{Q} \pm 2\sqrt{\frac{1}{Q}} \quad . \tag{2.15}$$

It is possible to compare this null distribution with that obtained from empirical asset returns. The remarkable result shown in [36] is that around 94% of empirical eigenvalues fall within the interval $[\lambda_- \ \lambda_+]$ and are therefore indistinguishable from noise. The 6% largest eigenvalues are instead well separated from this random bulk, and represent the informative part contained in the correlation matrix. In particular the largest one - λ_{max} - is typically one order of magnitude greater than the second largest.

These results are confirmed by our own analysis on the equity data set. We have $T = 4025$, $N = 342$ and $Q = T/N = 11.769$, that yields $\lambda_+ = 1.668$ and $\lambda_- = 0.502$. The corresponding theoretical spectrum for uncorrelated data is shown in Fig. 2.12 b) (solid curve); such spectrum is consistent with the empirical distribution of eigenvalues obtained from the correlation matrix of randomly shuffled log-returns (Fig. 2.12 b), dotted line), but it is remarkably different from the empirical spectrum obtained by the original correlation matrix $\{\rho_{ij}\}$ shown in Fig. 2.12 a). In particular, we have found that 14 eigenvalues (about 4% of the total) are greater than the upper bound λ_+, indicating significant correlation factors: together they account for the 48% of the total variance. The greatest eigenvalue, $\lambda_{max} = 108$, is almost ten times bigger than the second greatest (inset of Fig. 2.12 a)) ; it explains 30.01% of the total variance. These results are consistent with the existing literature [36].

The implications for portfolio optimization tools are especially relevant, since smallest eigenvalues are those that most affect the asset allocation according to the Markowitz method [36, 146, 147]. At the same time they corroborate Capital Asset Pricing Model (CAPM) and Arbitrage Pricing Theory (APT) [122] approaches to pricing, since these models assume a limited number of common factors as source of dependencies among different assets. However, it is worth mentioning that the interpretation of the empirical bulk of small eigenvalues is debated [32, 148, 149]: for example in [150] the authors have shown that such bulk can also be produced by the superposition of small structures of significant dependence [150].

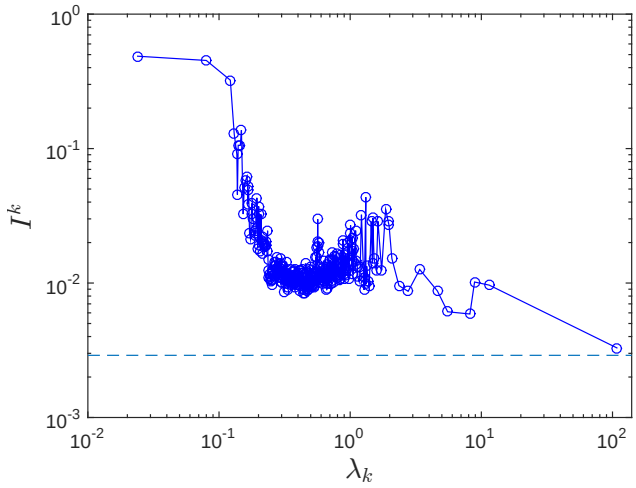

Fig. 2.13 **Contribution of the assets to each eigenvector**. Inverse participation ratio I_k against eigenvalue λ_k for each stock, $k = 1,...,N$. The horizontal dotted line represents $1/N$, the lowest possible value of IPR, when all assets participates equally to the eigenvector. As one can see, the lowest I_k is quite close to the horizontal line and corresponds to λ_{max}; in this sense λ_{max} represents the average market.

The financial meaning of the largest eigenvalues becomes clearer when the eigenvector are analysed. It is indeed possible to quantify the contribution of an asset i to an eigenvector u^k by looking at its component u_i^k: the larger its absolute value is, the stronger the asset contribution is. An overall participation degree for the eigenvector u^k can be computed by means of the Inverse Participation Ratio (IPR) [32]:

$$I^k = \sum_l^N [u_l^k]^4 \quad , \tag{2.16}$$

where u_l^k , $l = 1,...,N$ are the components of eigenvector u^k. I^k ranges between $1/N$ (when all assets participates equally, $u_l^k = 1/\sqrt{(N)}$ for each l) and 1 (when only one asset contributes to u^k). We have calculated the IPR for each eigenvalue in the data set; in Fig. 2.13 we show I^k as a function of eigenvalue λ_k. As we can see, the largest eigenvalue λ_{max} has the lowest IPR, equal to 0.0033, very close to the lower bound $1/N = 0.0029$ (dotted horizontal line). It therefore displays a quite homogeneous contribution from all assets and it can be interpreted as the overall market

Table 2.2 **Summary table of ρ_{ij}^R statistics**. We report the main sample features for the Pearson coefficient computed on log-returns detrended of market mode (first row) and on randomly shuffled log-returns detrended of the market mode (second row).

Type	Mean	Standard Deviation	Skewness	Min	Max
detrended log-returns	-8.9951e-04	0.0787	2.6504	-0.1938	0.9562
shuffled detrended log-returns	-0.0027	0.0158	0.0035	-0.0819	0.0855

influence [32, 36], that we call hereafter "market mode". Other large eigenvalues display significant contributions from specific categories of assets, such highly capitalized stocks and industrial sectors [32].

2.4.3 Subtracting the market mode

The predominance of λ_{max} is so strong that we may wonder whether the influence of the market mode conceals other levels of interactions among the assets. To answer this question it has been suggested to analyse the correlation matrix of detrended log-returns, i.e. log-returns subtracted of the average return over all the stocks [32, 68] $r_M(t) = 1/T \sum_i r_i(t)$. Specifically, following [68], we have considered a single factor model for each stock i:

$$r_i(t) = \alpha_i + \beta_i r_M(t) + c_i(t) \ , \tag{2.17}$$

where the residuals $c_i(t)$ are the log-returns detrended by the market mode. After estimating the coefficients α_i and β_i, the residuals $c_i(t)$ can be calculated and used to evaluate the new correlation matrix [68]. We denote this matrix with $\{\rho_{ij}^R(T)\}$. We refer to the analyses based on this kind of correlation matrix as the “detrended case”. These detrended correlation matrices are worth analysing since they have been found to provide a richer and more robust dependence structure [68] that can carry information not evident in the original correlation matrix [120]. In Tab. 2.2 and Fig. 2.11 b) the main features of $\{\rho_{ij}^R(T)\}$ are shown, compared with the corresponding shuffled correlation. Although the average is now very close to zero, the standard deviation of $\rho^R(T)$ is still

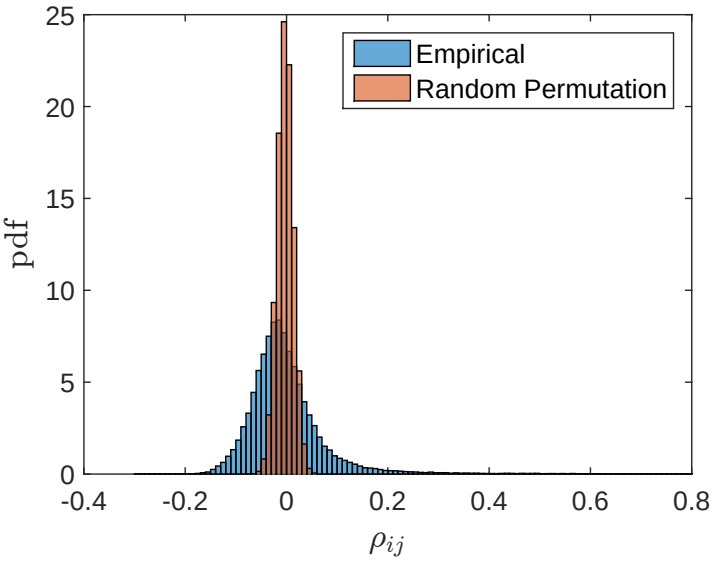

Fig. 2.14 **Correlation coefficients distributions for detrended log-returns**. Upper graph: histogram of correlation matrix entries $\rho_{ij}^R(T)$ detrended of the market mode, for empirical (blue) and shuffled log-returns (orange). Although the empirical mean is zero, empirical distribution is more spread and skewed towards positive values.

greater than the random case, and skewness is even greater than the non-detrended case in Tab. 2.1. This non-random structure is evident in terms of eigenvalues as well: in Fig. 2.15 the sample distribution of $\{\rho_{ij}^R(T)\}$ eigenvalues is shown, revealing that 29 eigenvalues (8.5% of the total) are greater than λ_+ and therefore carry meaningful information (they account for the 33.2% of total variance). Unlike the spectrum of $\{\rho_{ij}(T)\}$ though, there is no dominant eigenvalue. We can therefore conclude that the dependence structure remains very different from the random case also when the market influence is removed; besides the distribution of the remaining significant eigenvalues is more homogeneous in terms of total variance explained.

2.4.4 Dynamical evolution of correlation

For what concerns the temporal evolution, financial correlation displays significant changes over time [76, 119, 110, 121]. In order to study the dynamic of correlation, we

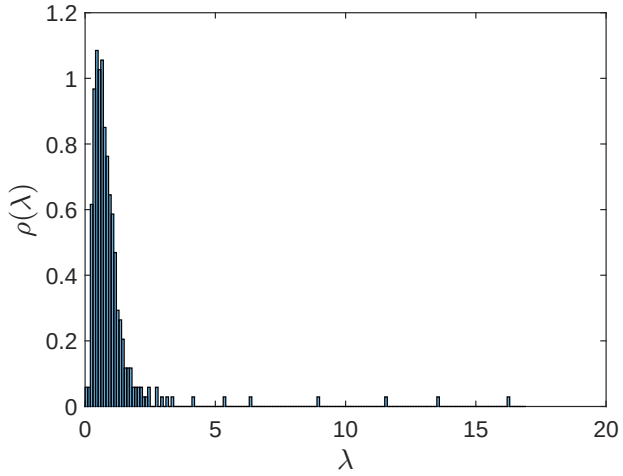

Fig. 2.15 **Correlation matrices spectra for detrended log-returns**. Eigenvalue distribution of correlation matrix $\{\rho^R_{ij}(T)\}$ from detrended log-returns. The distribution is still much different from the random one shown in Fig. 2.15 b).

need to modify the Pearson coefficient defined in Eq. 2.11. Indeed, the Eq. 2.11 assigns the same weight to each pair $\{r_i(t), r_j(t)\}$, regardless of t; however, observations at different times can have different relevance in a dynamic setting. Typically more recent observations carry more valuable information, especially in non-stationary scenarios that often occur in Finance. We therefore need a weighted version of Eq. 2.11. First of all let us define a set of moving time windows of length θ and shift dT between adjacent windows; in formula:

$$T_k = [t_{1+(k-1)dT}, t_{1+(k-1)dT+\theta}] \ , \tag{2.18}$$

with $k = [1, ..., n]$. On each time window we calculate a weighted version of the Pearson correlation matrix on log-returns [151, 113]:

$$\rho_{ij}(T_k) = \frac{\sum_{t=1}^{\theta} w_t (r_i(t) - \bar{r_i}^w)(r_j(t) - \bar{r_j}^w)}{\sqrt{\sum_{t=1}^{\theta} w_t (r_i(t) - \bar{r_i}^w)^2} \sqrt{\sum_{t=1}^{\theta} w_t (r_j(t) - \bar{r_j}^w)^2}} \ , \tag{2.19}$$

$$\bar{r}^w_{i/j} = \sum_t w_t r_{i/j}(t) \tag{2.20}$$

$$w_t = w_0 \exp\left(\frac{t-\theta}{T^*}\right) \quad , \tag{2.21}$$

where T^* is the weight characteristic time ($T^* > 0$) that controls the rate at which past observations lose importance in the correlation. For $T^* \to \infty$ the weighting becomes uniform (i.e. each observation has the same weight), whereas $T^* \to 0$ puts all the weight on the most recent observation. We have chosen $T^* = \theta/3$ according to previously established criteria [103]. w_0 is a constant connected to the normalisation constraint $\sum_{t=1}^{\theta} w_t = 1$.

From this set of correlation matrices $\{\rho_{ij}(T_k)\}$ we can derive a number of temporal measures. The first and simplest is the average correlation, defined as:

$$\langle \rho(T_k) \rangle_{ij} = \frac{2}{N(N-1)} \sum_{i \neq j} \rho(T_k)_{ij} \quad . \tag{2.22}$$

$\langle \rho(T_k) \rangle_{ij}$ is shown in Fig. 2.16 a) (blue circles) for $\theta = 1000$, $n = 100$ and $dT = 30$. As one can see, two main increases occur in correspondence with the 2002 recession and 2008 credit crunch. The increase of correlation in correspondence with financial crises is a well known fact [3, 121]. Interestingly the average correlation remains relatively high years after the outbreak of 2008 financial crisis, despite the average price resumes rising in 2009. This is consistent with the observed shift in investors behaviors after the crisis [152]. We have verified that such evolution is robust against change in window sizes, as shown in Fig. 2.16 b).

In Fig. 2.16 the average correlation for detrended correlation matrices, i.e. $\langle \rho^R(T_k) \rangle_{ij}$, is shown as well. As one can see, the subtraction of the market mode decreases by more than two orders of magnitude the average level of correlation, pointing out again the important role of the market factor in the dependence structure. However, we can still observe the increase correspondent to the financial crisis in 2007-2008. Moreover, and interestingly, the level of correlation reduces after a peak in 2009, unlike the non-detrended case. This fact suggests that, although the market mode plays an important

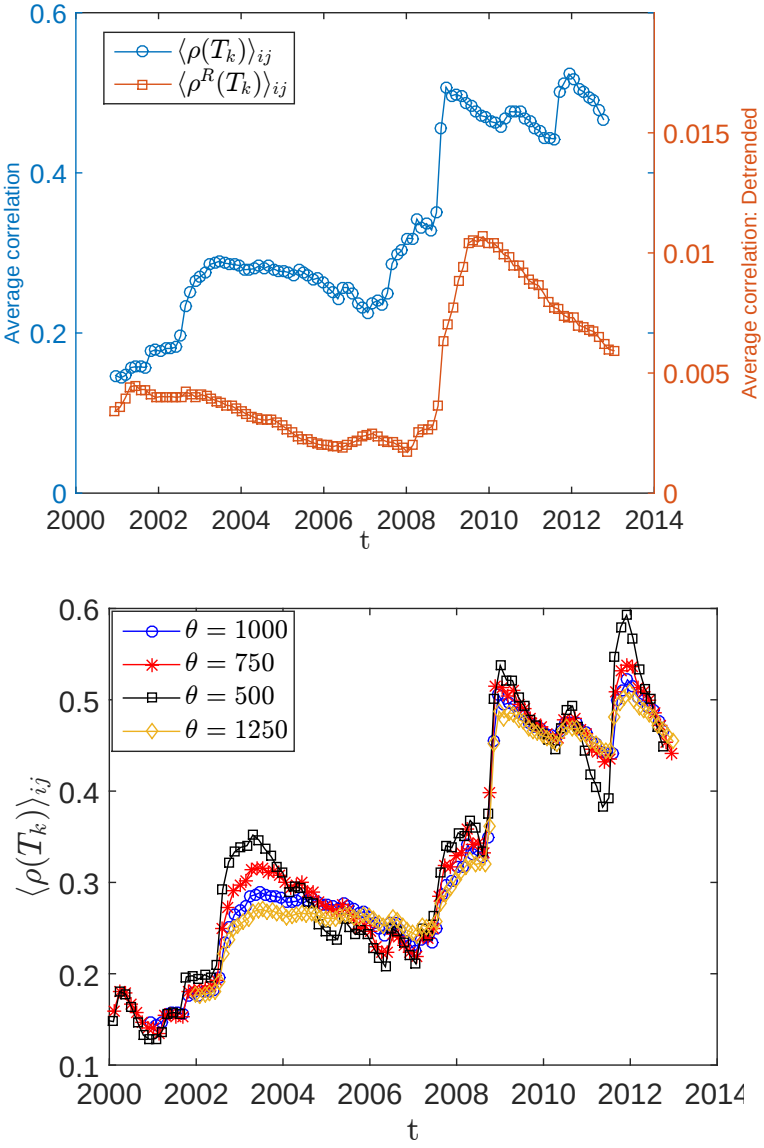

Fig. 2.16 **Dynamic evolution of average correlation**. The figure reports the average correlation for each time window T_k with $k = 1, ..., n$ ($n = 100$, each time window has length $L = 1000$ trading days), for both non-detrended (blue circles) and detrended log-returns (green squares). The average correlation is highly reduced by detrending the market mode.

role in terms of average amount of correlation, yet the peak of the last financial crisis seems not to be only a global market trend. We therefore suggest that it could involve, to some extent, the internal dynamics among stocks that remains after the subtraction.

The evolution of $\langle\rho(T_k)\rangle_{ij}$ suggests that assuming a stationary dependence structure is not reasonable. In fact non-stationarity of stocks correlation has been checked by

means of statistical tests under assumptions of normality [38], raising new doubts over the reliability of sample correlation for portfolio optimization. Non-stationarity makes the naive use of historical data even more questionable than the noise issue does, since it implies that longer time windows do not necessarily provide better estimates of correlation. Still some regularities can be observed in this dynamic picture. Evidences of mean reversion are found in correlation variations, as well as positive autocorrelation among correlation values [110].

Another important feature of financial correlation evolution is its relation with market returns: periods of large losses and negative returns are characterized by high correlations as well [110] [121]. This is especially true during financial crises, as we pointed out. Although often mentioned as a manifestation of non-stationarity, it has been shown that this empirical fact can be actually explained assuming non-linearity and tail-dependence within a stationary dependence structure [49] [47]. Such relation between average correlation and market returns seems to hold time-lagged as well [153], indicating that current market returns carry information on future average correlation (but not vice-versa).

2.4.5 Economy-related information

The dependence structure of assets returns contains a significant amount of economy-related information. Correlation matrices of equities display a hierarchical structure that partially mirrors the industrial sector partition of the stocks [118] [53], with higher correlation among equities belonging to the same sector; this is more evident with correlation-based networks than with spectral methods [154]. Assets tend to group according to geographic membership too, although this influence is weaker [155] and emerges only when the industrial sector impact is subtracted [120]. However this is quite a recent effect of globalisation, as before nineties pure country factors seemed to dominate global industrial factors [156, 157]. Moreover, the irreversibility of this effect is debated [158].

The similarity between dependence structure and industrial sectors is highly time-dependent [70], decreasing and almost disappearing during periods of financial crises [76]. When correlation across different asset classes is taken into account (bonds, equities, currencies, commodities) an analogous picture emerges: spectral analyses [119] show how main risk factors can be identified with each asset class, although again this structure is strongly time-dependent and breaks during turbulent periods. For instance during the 2007-08 financial crisis all asset classes but bonds became highly correlated among each other, regardless of the asset class, while bonds formed a separate cluster of strongly correlated assets [119]. This behavior is known as flight-to-quality and is caused by investors selling risky assets and buying bonds during times of negative returns [48].

2.4.6 Time scale

Financial correlations display peculiar behaviors when the sampling frequency τ changes. In particular correlation among assets decreases when sampling frequency increases: this phenomenon is known since 1979 and is called Epps Effect [159]. Different explanations have been suggested for this empirical fact. According to [159], Epps Effect is due to asynchronous trading between correlated stocks, caused for instance by not instantaneous flow of new information among traders; this would be consistent with the finding of lagged correlations between stocks. Another work [160] indicates also discretization effects (that is, effects linked to the discrete nature of price changes) as a possible cause. In [161] herding behavior is suggested as a further reason, whereas in [162] the stress is on the different time horizons among the market participants. Another phenomenon connected to the sampling frequency is described in [68], where it is shown how the similarity between dependence structure and industrial sector partition decreases when Δt increases. This means that economic-related information emerges from dependencies only above a certain sampling frequency. Interestingly, when either the first eigenvector or the average market return is subtracted this effect disappears

and industrial sector structure becomes visible at even 5 minutes of sampling frequency [68]: the market driving factor is therefore responsible for hiding such structure at high sampling frequencies.

2.4.7 The limit of Pearson coefficient: non-linearity in financial correlation

Despite its popularity, Pearson coefficient has some important pitfalls as a measure of dependence [51]. In particular its connection to linear regression makes it unsuitable for pairs of returns r_i and r_j whose relation is not linear. A well-known example is the case $r_j = r_i^2$ for $r_i \sim N(0,1)$: even if there is a deterministic dependence between r_i and r_j, it turns out that $\rho_{ij} = 0$. In general, uncorrelation does not necessarily imply independence, unless r_i and r_j are drawn from an elliptic distribution [51]. As a consequence of its linear nature, Pearson coefficient is not invariant under non-linear transformations of r_i and r_j, which is another undesirable feature.

In fact a number of empirical results have found that Pearson coefficient only is not sufficient for assessing the dependence between financial assets. In particular a certain degree of non-linearity emerges in form of exceedance correlation [49, 50], that is the Pearson correlation computed only on returns above/below a certain threshold quantile [113]. Under multivariate normal assumptions we would expect exceedance correlation to go to zero when the threshold selects returns farther and farther from the mean; empirically, equity data show instead increasing exceedance correlation for negative tail (but not for positive tail), as shown in [49, 50]. We will discuss non-linear measures of dependence, alternative to Pearson coefficient and better at capturing these features, in Chapter 7.

2.5 Summary

In this chapter we have reviewed the main empirical properties of financial time series and financial correlation. To this end we have performed a set of original analyses on a data set of stock prices, which covers a period of 15 years including the Dotcom bubble and the latest financial crisis of 2007-2008. What emerges is a quite complex picture, where high level of noise comes together with meaningful information. In particular the correlation matrix on log-returns turns out to be significantly different from what expected from uncorrelated processes; yet much statistical noise is present, with about 96% of eigenvalues falling within the bulk of the random model. The average correlation displays quite a dynamic behaviour, in particular with a steep increase during the financial crisis, as already observed in the literature [3, 121]. If such average trend is removed we find a structure that is still meaningful compared to the random model, and notably some patterns due to the financial crisis is still observable.

As our analyses confirm, a major challenge for any attempt to deal with financial dependencies is their non-stationarity. Such lack of stability invalidates many traditional econometric tools designed to deal with risk. In the rest of this thesis we aim at addressing these issues and others, by proposing a set of analyses that address naturally the problem of non-stationarity and allow to compare different dependence measures at the same time. These analyses rely on an information filtering technique quite popular in Econophysics, namely the correlation-based filtered networks. In the next chapter we introduce such methods and we summarise the main empirical insights that its application to financial data has provided so far.

Chapter 3

Correlation-based filtered networks

In this chapter we introduce different filtering techniques that we will use extensively throughout this thesis, namely correlation-based filtered networks (also called simply correlation-based networks in this thesis). In particular, we review the main types of correlation-based networks, such as Minimum Spanning Tree and Planar Maximally Filtered Graphs. We apply such tools to the equity data set and we discuss the insights they provide into the dependence structure evolution. Part of the results and analyses presented in this chapter has been published in the paper "Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Method" in 2015 [109].

3.1 Introduction

The Pearson coefficient is an inherently pairwise measure which does not have a natural extension to more than two variables. This limitation concerns the applicability of significance tests too. Hence this measure does not allow to assess the degree of significance of the overall dependence structure when the number of assets N is much higher than 2. One could apply for instance the statistical test of Eq. 2.13 for each one of the $N(N-1)/2$ pairs of assets, but this analysis would be only a collection of pairwise tests discarding completely potential information regarding the global system

of interactions, such as common factors influencing more than one asset. We therefore need tools that are conceived to deal with complex systems of many interacting nodes, able to unveil and take into account the overall structure of interactions. An example of such tools are the so-called "dimensionality reduction techniques" [15], which provide a simplified representation of the original system by exploiting its patterns and regularities; for this reason they are valuable for data visualization as well. An example is the Random Matrix Theory [33, 35], discussed in Chapter 2, and the related Principal Component Analysis [163].

The techniques we will use in this thesis are known as correlation-based filtered networks. These tools exploit Network Theory [12, 11, 13] to filter information from any kind of distance or similarity matrix [56, 55, 164]; when applied to correlation matrices they can reduce statistical noise and unveil their hidden hierarchical structures. They have been originally introduced in the context of optimization of electrical networks [165] in the 20's, and then applied in Finance for the first time by Mantegna at the end of 90's [53]: since then they have been used extensively in the Econophysics literature, applied on a large variety of financial market data [53–60].

We review the main types of correlation-based networks, that is the Minimum Spanning Tree [53, 65], the Asset Graph [54] and the Embedded Graph [164]; among the Embedded Graphs we will focus on Planar Maximally Filtered Graph [55, 56]. Furthermore, we discuss how these tools are deeply linked to hierarchical clustering methods; in particular we introduce the Directed Bubble Hierarchical Tree method [66], a recently introduced clustering method based on Planar Maximally Filtered Graphs, that in this thesis will be applied for the first time to financial data (see Chapter 4). Moreover, we demonstrate the power of network filtering by analysing correlation-based networks constructed from the equity data set: we introduce and apply network theory concepts such as degree distribution and degree-degree correlation.

A review of the main empirical findings that correlation-based networks made possible is presented in this chapter as well. In particular, we discuss contributions

concerning the economic information contained in the dependence structure [53, 120, 70], its degree of non-stationarity [73, 75] and its response to major events affecting financial markets, such as financial crises and news [76, 77]. Moreover, we show how these tools can have practical applications in enhancing portfolio optimisation methods [79, 81].

The structure of this chapter is as follows. In Section 3.2 we introduce the concept of correlation-based network applied to financial data, by reviewing the main types of tools introduced in literature. Furthermore, we apply these tools to the equity data set, to demonstrate their filtering power and analyse the networks structure. In Section 3.3 we review the main empirical findings and applications that the use of correlation-based networks made possible. In Section 3.4 we describe the main hierarchical clustering methods and we discuss their connections with filtered correlation-based networks.

3.2 Financial network: definitions

Correlation-based networks are sparse network representations of dependence matrices (not necessarily correlation), obtained by interpreting these matrices as adjacency matrices [13] and then performing some sparsification algorithm. The purpose is to filter most of the statistical noise and redundancy, while retaining the backbone of the dependence structure. The sparsification algorithm is what distinguishes different types of correlation-based networks. In this section we review the main types introduced in literature. In the following, we refer to the original, unfiltered $N \times N$ dependence matrix as S. In the terminology of clustering methods, S is the similarity matrix (as it measures how "close" the N elements are to each other), from which a correspondent distance matrix D can be calculated. It can be shown that an appropriate distance matrix, when the dependence matrix is Pearson correlation (i.e. $S_{ij} = \rho_{ij}$), is [53]:

$$D_{ij} = \sqrt{2(1 - S_{ij})} \ . \tag{3.1}$$

The advantage of this choice is that this D fulfills the three properties of a metric distance:

$$D_{ij} = 0 \;\Leftrightarrow\; i = j, \tag{3.2}$$

$$D_{ij} = D_{ji}, \tag{3.3}$$

$$D_{ij} \leq D_{ik} + D_{kj} \tag{3.4}$$

Moreover the distance D in Eq. 3.1 has a natural interpretation, since D_{ij} equals the Euclidean distance calculated between standardized returns of asset i and j [3].

The input of a sparsification algorithm can be either the similarity or the distance matrix, and the algorithm description changes accordingly; to be consistent with notation in [53], we here consider the distance matrix as input matrix.

3.2.1 Minimum Spanning Tree

In Network Theory, a tree is defined as a graph with no loops. Given a connected network on a set of objects, the Minimum Spanning Tree (MST) is the tree connecting all the objects and minimizing the sum of weights [11]. It is a powerful tool that has been used mainly in optimization problems such as telecommunication networks, electrical grids and water supply networks [167]. One of the first application of this tool dates back to 20's, with the purpose of designing an efficient electrical network [165]. Other applications of MST include taxonomy [65] and cluster analysis [62]. The MST has been the first tool from Network Theory to be applied to financial data [53].

Some important properties hold for a MST. If the number of nodes is N, the total number of edges in a MST is $E = N - 1$. Moreover, if all edges have different weights, it can be shown that only a MST exists given the original network [11]. In Fig. 3.1

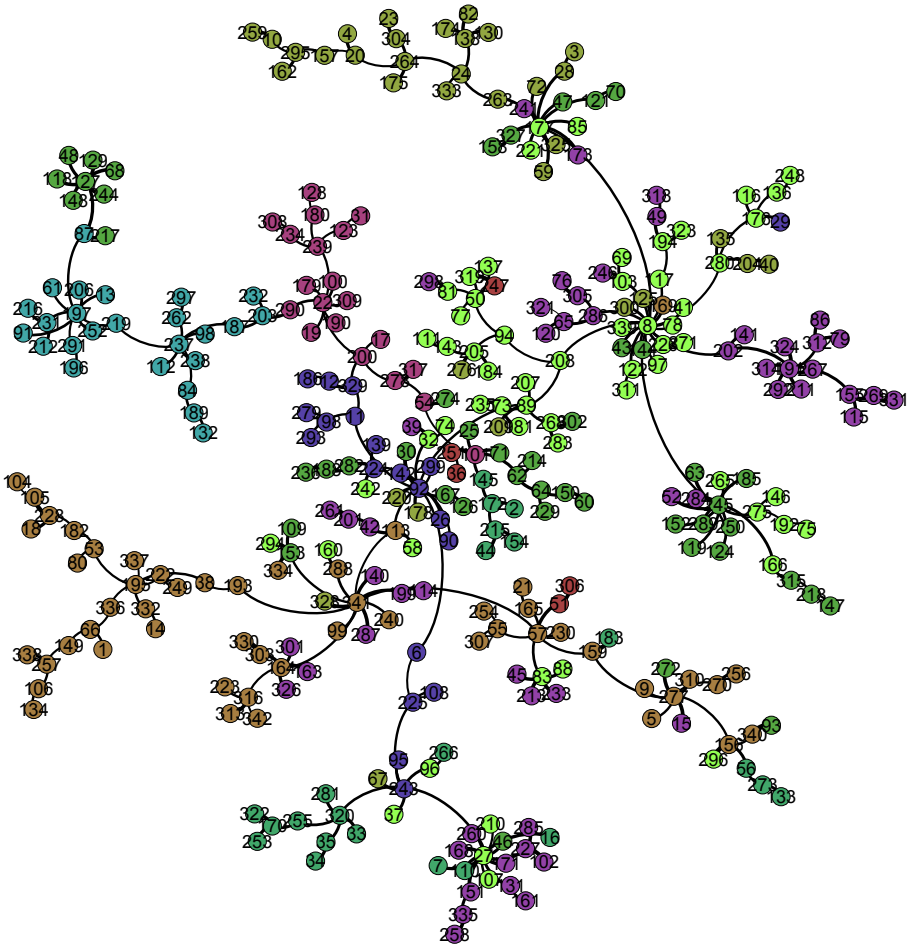

Fig. 3.1 **MST from Pearson correlation among 342 US stocks**. We have built the MST from the correlation matrix $\{\rho_{ij}\}$ which we have computed on the data set of 342 US stocks, over a 15 years time window from 02/01/1997 to 31/12/2012. Different colors identify different industrial sectors (ICB classification). Visualisation elaborated with Gephi [166].

we show the MST which we have computed from the data set correlation matrix $\{\rho_{ij}\}$ introduced in Chapter 2. In the picture, colours represent different ICB industries. One

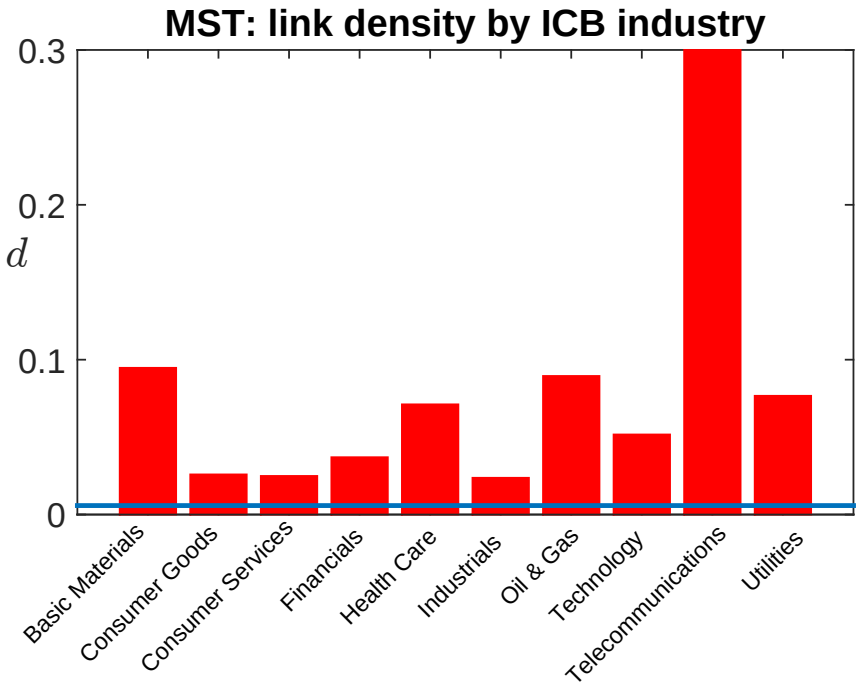

Fig. 3.2 **Economic information extracted by the MST topology**. Density of links in our MST computed within each ICB industry, compared to the average density in the network (horizontal blue line). All industries display an internal connectivity greater than the average, indicating that the network filtering is unveiling a meaningful structure.

can observe how stocks from the same industry tend to gather together in the network, as observed for the first time in [53]. The visual intuition is supported by a quantitative test: we have calculated the average density of links in the network (defined as the ratio between the number of links E and the total number of pairs of nodes, i.e. $N(N-1)/2$ [12]) as well as the density of links within each industry, and we have found that the intra-industry density is always greater than the average density. As shown in Fig. 3.2 the industries with higher density are Telecommunications, Basic Materials and Oil & Gas. This result demonstrates the power of network filtering: a meaningful structure in terms of industrial activity emerges from the data thanks to the sparsification algorithm of MST.

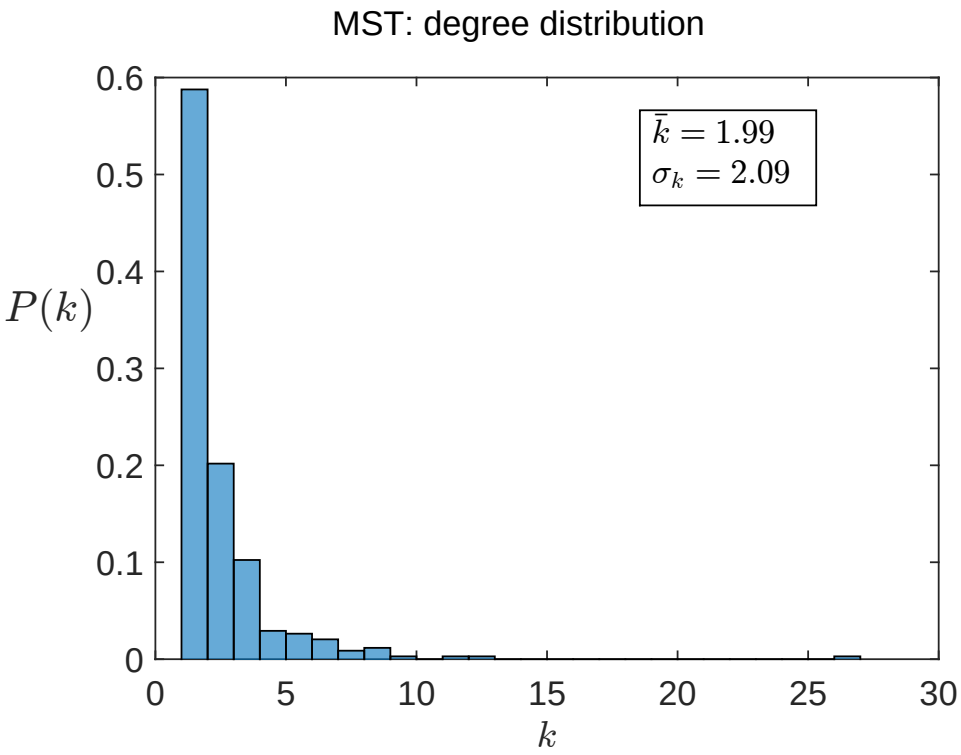

Fig. 3.3 **Degree distribution for the MST**. Histogram of degrees calculated from our MST. The distribution shows a quite broad range of degree values, with a maximum degree of 26.

Once we have mapped the original correlation matrix into a network, we can exploit the tools of Network Theory to gain an insight into the dependence structure. Let us denote with $\{a_{ij}\}$ the adjacency matrix [11] of the network, such that $a_{ij} = 1$ if and only if there is a link between nodes i and j. Given the adjacency matrix, the simplest network analysis that we can perform concerns the statistics of degrees, namely the number of connections that each nodes has in the network [12]. We define the degree of node i as $k_i = \sum_j a_{ij}$; we denote the network degree distribution with $P(k)$. In our MST we have found that $P(k)$ covers a broad range of values, with k ranging from 1 to 26; the mean degree is 1.99 and standard deviation 2.09 (see Fig. 3.3 a)). The number of points in the tail of $P(k)$ is not sufficient to infer the exact decay function. However, in

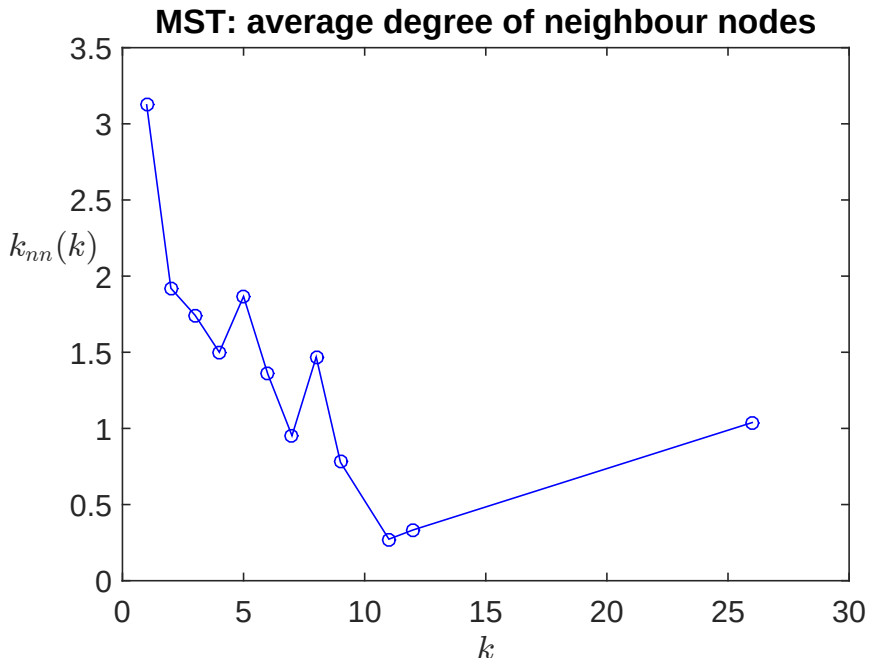

Fig. 3.4 **Average neighbour degree as a function of node's degree in the MST**. The average neighbour degree $k_{nn}(k)$ from our MST, as a function of node's degree k. There is an overall decreasing trend which indicates that the MST is a disassortative network: high degree nodes tend to be connected with low degree nodes, and vice-versa.

[168] has been shown that MST on financial data displays power-law behavior in the tail of the degree distribution, a typical signature of complexity in networks [13, 12].

A deeper insight into the topology is provided by the conditional distribution $P(k'|k)$, that quantifies the probability of selecting a node with degree k' among the neighbours of a node with degree k [12]. Since the estimation of $P(k'|k)$ is quite problematic in sparse networks with fat-tailed $P(k)$ [12], it is common to analyse the derived quantity $k_{nn}(k)$ defined as $k_{nn} = \sum_{k'} k' P(k'|k)$ [169]. $k_{nn}(k)$ represents the average degree of a node with degree k. If $k_{nn}(k)$ increases with k it means that high degree nodes (hubs) tend to connect with other high degree nodes and the network is called "assortative" [13]; whereas if $k_{nn}(k)$ decreases with k, high degree nodes tend to connect with low degree nodes and the network is called "disassortative" [13]. We have calculated $k_{nn}(k)$ for our MST and we show in Fig. 3.4 the resulting function; as one can see the trend is overall decreasing (a part from the point with the highest degree, which is however

less reliable as it includes observations from only one node), therefore the network is disassortative. Hence the network of dependence, as shown by the MST, is characterised by hubs of assets highly connected with poorly connected assets.

Given the original distance matrix D, several algorithms can be run to calculate the correspondent MST. Among the most common, we here recall the Kruskal's [170], the Borůvka's [165], the Prim's [171] and the reverse-delete algorithms [170]. For all of them, the computational time is $O(E\log(N))$. More recent algorithms are able to find the MST in linear computational time with respect to E, that is $O(E)$ [172, 173].

3.2.2 Asset Graph

The MST is obtained by imposing a topological constraint (tree structure) along with the minimization of weights. If these two constraints are replaced by a condition on a maximum acceptable distance D_{max} - that is, no distance greater than D_{max} can be added to the graph - the resulting network is a so-called threshold network, or Asset Graph (AG) in Econophysics literature [54, 71]. Unlike the MST, the Asset Graph is not necessarily connected. In particular, for networks constructed from financial data, removing links with weaker correlation makes the network disconnect relatively soon (with about 30% of links removed in [174]) and sooner than removing strong correlation links, implying that the former contribute to the overall connectivity more than the latter. Strong correlation links on the other hand tend to contribute more to intra-clusters and intra-sectors cohesion [174]. These networks exhibit power-law tails for both the degree and the strength (i.e. the sum of distances over the links incident to each node) distributions, for a broad range of D_{max} [175]; in particular for the strength distribution such scale-free behaviour holds even in the limit of fully connected graph, that is $D_{max} = 1$ [176]. In [177] the authors show that the way in which an Asset Graph changes with the threshold is sensitively different from a random graph, displaying in particular a much higher clustering coefficient.

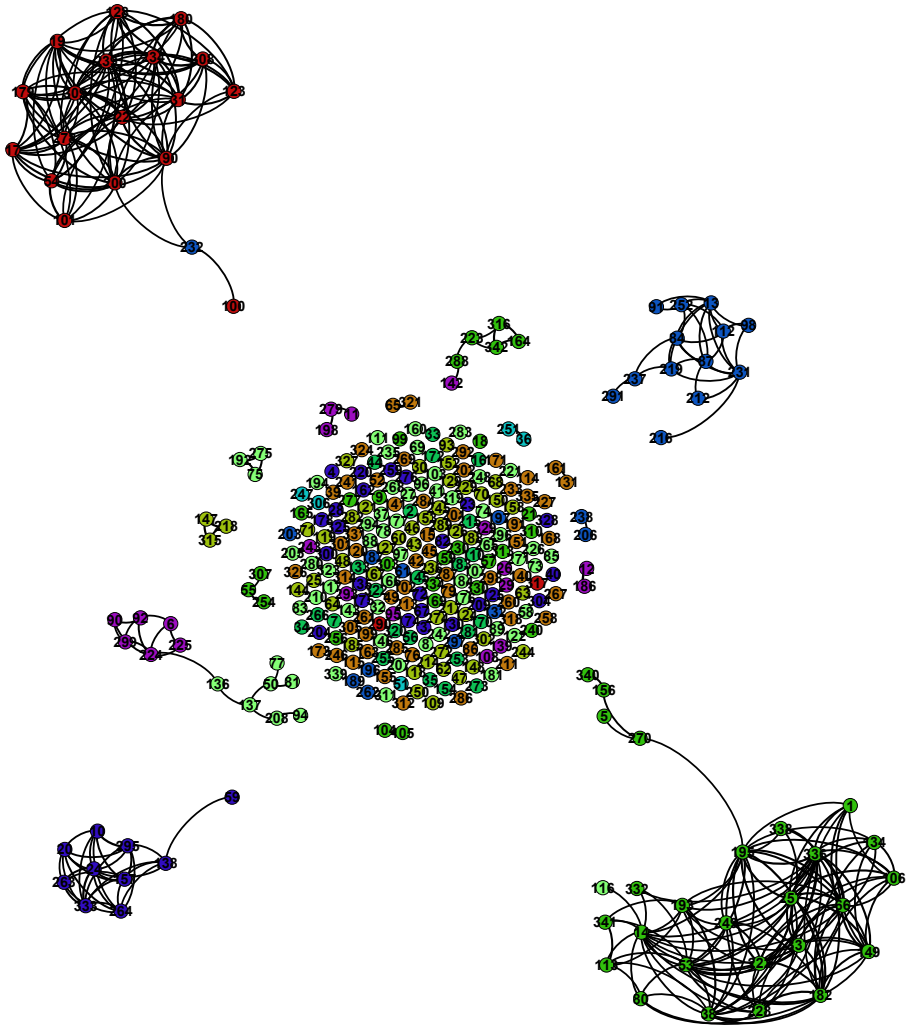

Fig. 3.5 **Asset Graph on Pearson correlation among 342 US stocks**. We have built the AG from the correlation matrix $\{\rho_{ij}\}$ which we have computed on the data set of 342 US stocks, over a 15 years time window from 02/01/1997 to 31/12/2012. Different colors identify different industrial sectors (ICB classification). Visualisation elaborated with Gephi [166].

The threshold requirement on D_{ij} can be replaced by a requirement on the overall number E of edges in the graph; in [54] $E = N - 1$ was chosen, in order to have a

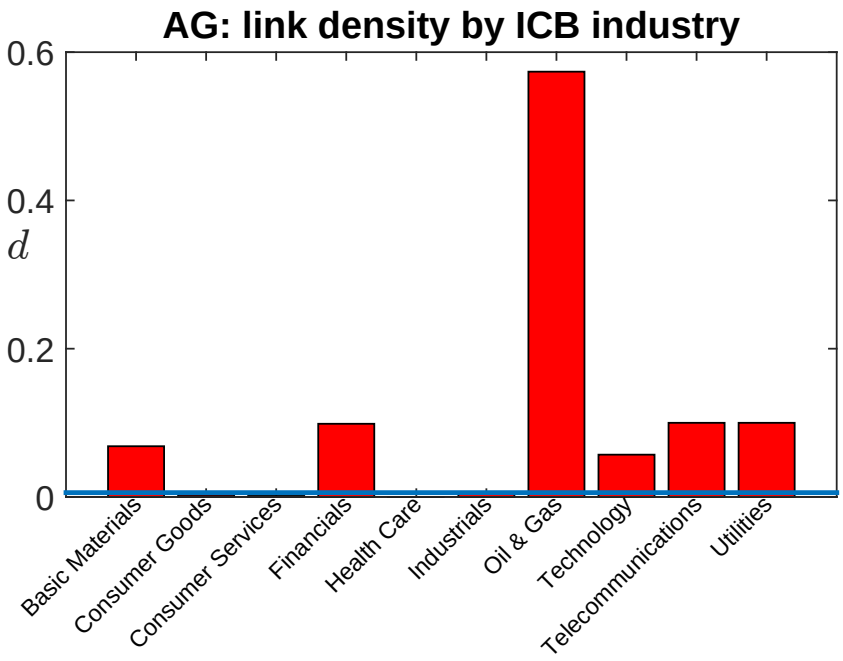

Fig. 3.6 **Economic information extracted by the AG topology**. Density of links in our AG computed within each ICB industry, compared to the average density in the network (horizontal blue line). All industries display an internal connectivity greater than the average, indicating that the network filtering is unveiling a meaningful structure.

network comparable with the MST. We have followed this approach and constructed an AG from our correlation matrix $\{\rho_{ij}\}$. The AG is shown in Fig. 3.5. As one can see, the network is disconnected: 230 nodes (about 67% of the total) have no connections at all. The nodes which are connected form compact and dense clusters, whose industrial composition is quite homogeneous, consistently with [174]. The density analysis confirms the picture: as we can see in Fig. 3.6, 6 ICB industries exhibit higher density than the average. In particular Oil & Gas is by far the most interconnected industry. Overall this structure indicates a strong heterogeneity in the correlation matrix: high correlation tends to occur mainly among stocks of certain industries. The degree distribution $P(k)$ is again fat-tailed, with a standard deviation of 4.17 (Fig. 3.7). In terms of degree-degree correlation the AG is assortative: as one can see from Fig. 3.8, the average neighbour degree $k_{nn}(k)$ is an increasing function of k. This is consistent

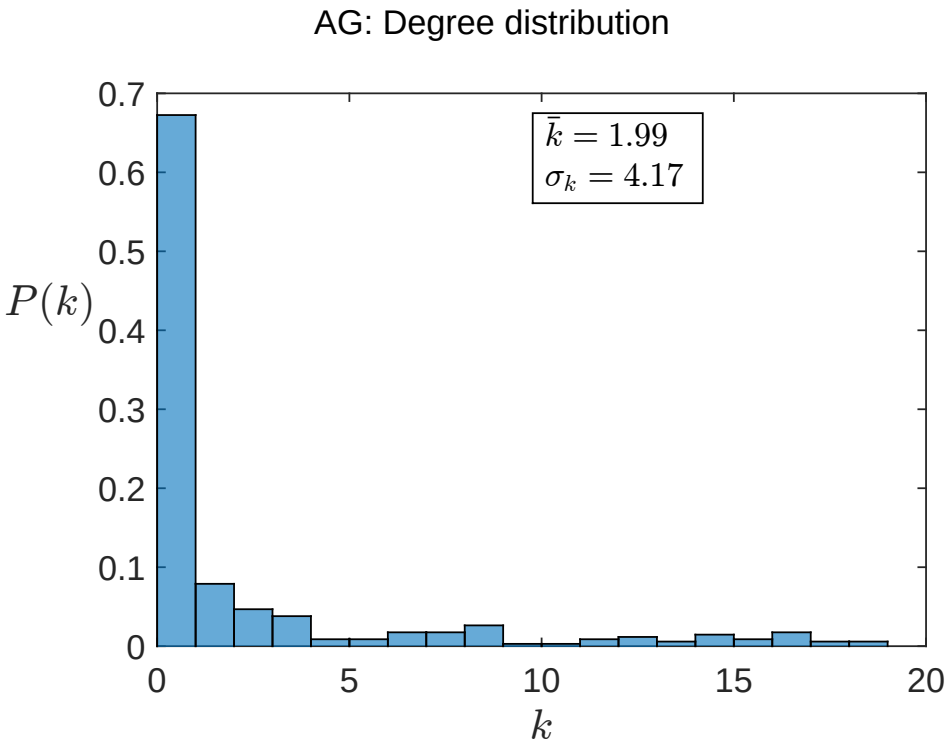

Fig. 3.7 **Degree distribution for the AG**. Histogram of degrees calculated from our AG. The distribution is again fat-tailed, with a standard deviation of 4.17 which is more than twice the MST standard deviation.

with the clustered structure we can observe from Fig. 3.5, where highly connected nodes are mostly connected with each other.

In [54] has been shown that the AG are less affected by non-significant, low correlations that are instead often kept by the MST. As a result the AG is more robust against time [54]. On the other hand, the MST, retaining both high and low correlations, is more able to uncover global, multi-scale structures of interaction. Indeed, in financial - and complex systems in general, several length scales coexist and thresholding at a given value introduces artificially a characteristic size that might hide effects occurring at other scales [178].

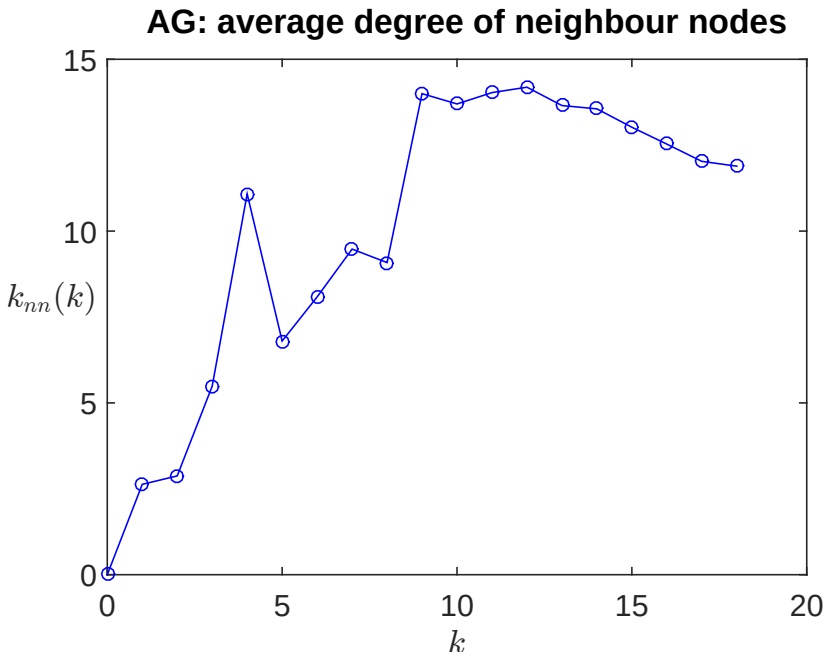

Fig. 3.8 **Average neighbour degree as a function of node's degree in the AG**. The average neighbour degree $k_{nn}(k)$ from our AG, as a function of node's degree k. The overall trend is now increasing, which indicates that the AG is an assortative network: high degree nodes tend to be connected with high degree nodes, whereas low degree nodes tend to be connected with other low degree nodes.

3.2.3 Embedded Graphs

As we have seen, the MST is constructed by imposing a topological constraint, namely the absence of loops. We can think of a more general class of topological constraints by means of the concept of "embedding": a graph can be embedded on a surface if it can be drawn on that surface without link crossing. In order for this embedding to be possible, it turns out that only a feature is relevant: the surface "genus" [55].

The genus g of a surface is the largest number of non-intersecting simple closed cuts that can be made on the surface without disconnecting a portion (equal to the number of handles in the surface). Intuitively, the higher the genus, the more handles are in the surface. Hence, the higher the genus the more links can be drawn on the surface without crossing, and then more networks can be embedded on that surface. In order to embed a fully connected graph of N nodes, a surface with genus $g \geq g_{max}$ is needed,

where $g_{max} = \lceil \frac{(N-3)(N-4)}{12} \rceil$ (where $\lceil x \rceil$ is the ceiling function that returns the smallest integer bigger or equal than x). This provides a topological criterion for filtering, which resembles the MST but is more flexible: namely, to reduce redundancy in the fully connected graph we can extract a connected subgraph which minimises the sum of weights/distances and which can embedded on surface with $g < g_{max}$.

In [164] ensembles of embedded networks have been analysed for a wide range of g, finding that networks with higher g show power-law degree distributions and small-world topology. The concept of embedding on surfaces provides therefore a quantitative way to tune the degree of information filtering by means of a single parameter, g, linking correlation-based networks to algebraic geometry [164].

Planar Maximally Filtered Graph

Among the topological constraints one can impose through the concept of genus, planarity is especially important. A graph is planar if it can be embedded on a plane without link crossing [55]: all trees are planar, but not vice versa. Requiring that a network be planar is equivalent to requiring that the network can be embedded on a surface with $g = 0$ (i.e., no handles, a sphere in topological terms). Therefore, all embedded graphs with $g = 0$ are planar; we call such graphs Planar Maximally Filtered Graphs (PMFG) [55]. Given the original dependence matrix, the PMFG is therefore that graph which minimises the sum of weights and is planar. The PMFG can be seen as a generalization of the MST, that is able to retain a higher amount of information [56, 179], having a less strict topology constraint allowing to keep a larger number of links. Moreover, the MST is a subgraph of PMFG [64].

The PMFG displays several properties. Similarly to the MST, if all edges have different weights there is a unique PMFG [56]. Each PMFG with N nodes contains exactly $E = 3(N-2)$ edges. Each node in a PMFG participates at least to one three-clique, that is a group of three nodes which are all connected to each other: hence the PMFG can be viewed as a triangulation of the sphere. Moreover, no clique of order

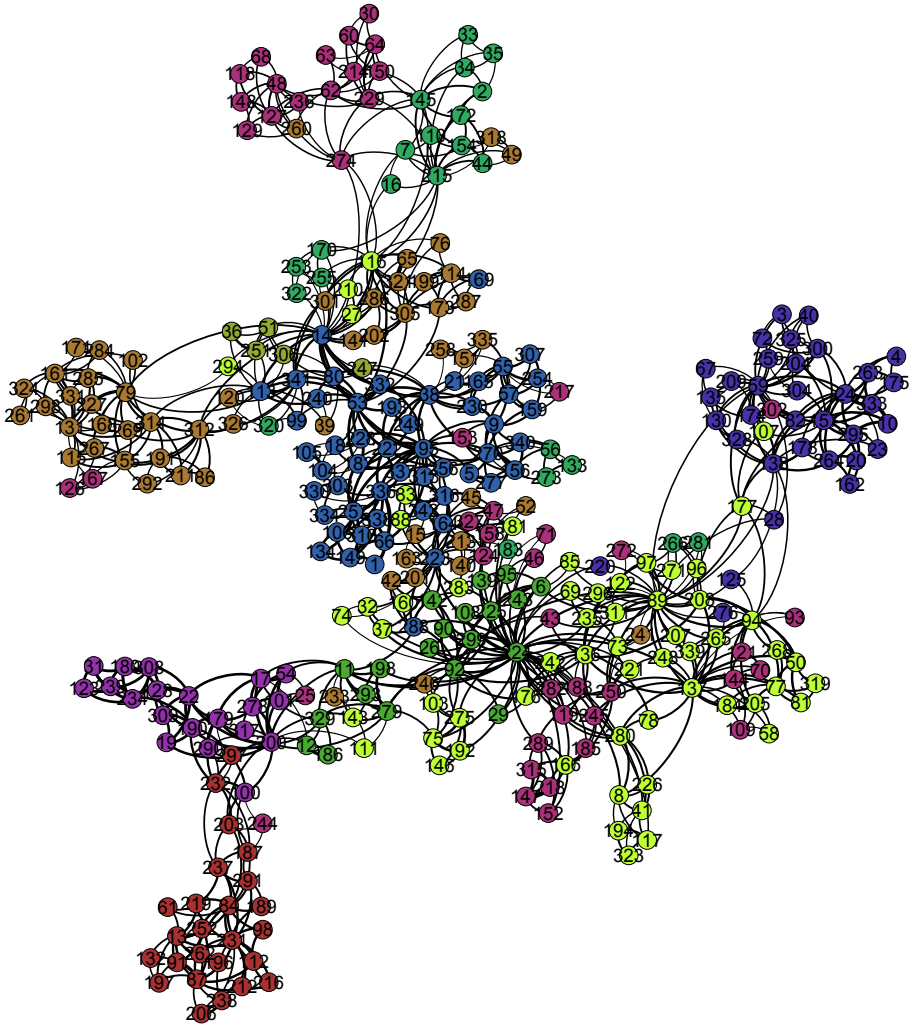

Fig. 3.9 **Planar Maximally Filtered Graph on Pearson correlation among 342 US stocks**. We have built the PMFG from the correlation matrix $\{\rho_{ij}\}$ which we have computed on the data set of 342 US stocks, over a 15 years time window from 02/01/1997 to 31/12/2012. Different colors identify different industrial sectors (ICB classification). Visualisation elaborated with Gephi [166].

greater than 4 can exist on a PMFG, a result known as Kuratowski theorem [180]. In terms of computational complexity the algorithm that builds PMFG is $O(N^3)$; recently

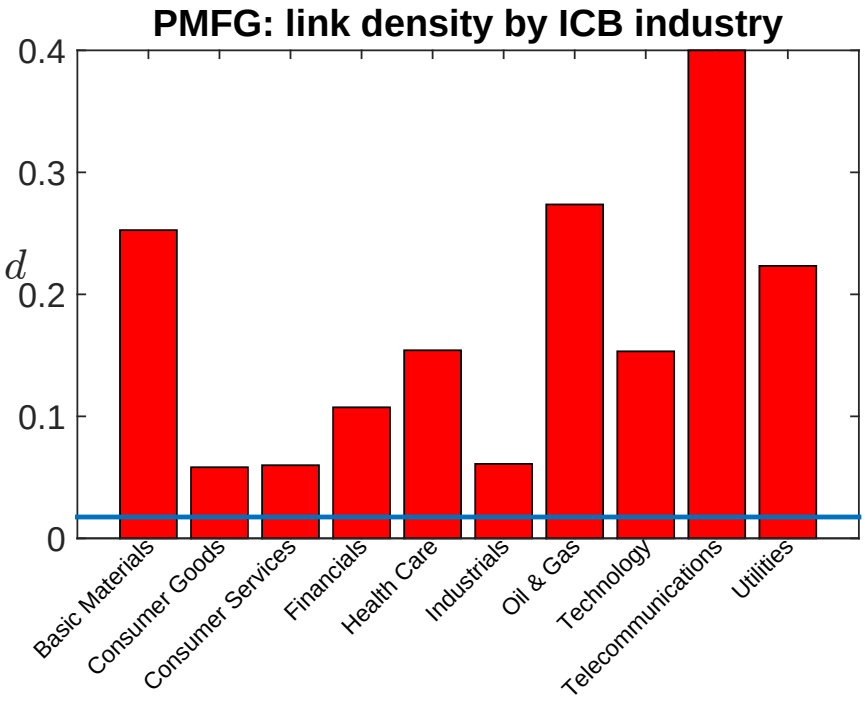

Fig. 3.10 **Economic information extracted by the PMFG topology**. Density of links in our PMFG computed within each ICB industry, compared to the average density in the network (horizontal blue line). All industries display an internal connectivity greater than the average, indicating that the network filtering is unveiling a meaningful structure.

[181] a new algorithm has been proposed, able to build an approximation to the PMFG (called Triangulated Maximally Filtered Graph, TMFG) with an execution time $O(N^2)$, making possible a much higher scalability and the application to Big Data [181].

We have computed the PMFG associated to the correlation matrix of the data set; the result is shown in Fig. 3.9. The structure resembles the MST topology, except for the larger number of links. The industry-related information, as quantified by the density of links, is indeed similar to the MST: as shown in Fig. 3.10, the most interconnected industry is again Telecommunications. However Utilites, Oil & Gas and Basic Materials emerge more clearly than in the MST case (Fig. 3.2). The topological properites are similar to those of MST. The degree distribution is again fat-tailed, as shown in Fig. 3.11, with largest degree equal to 55. In terms of degree-degree correlation, the average

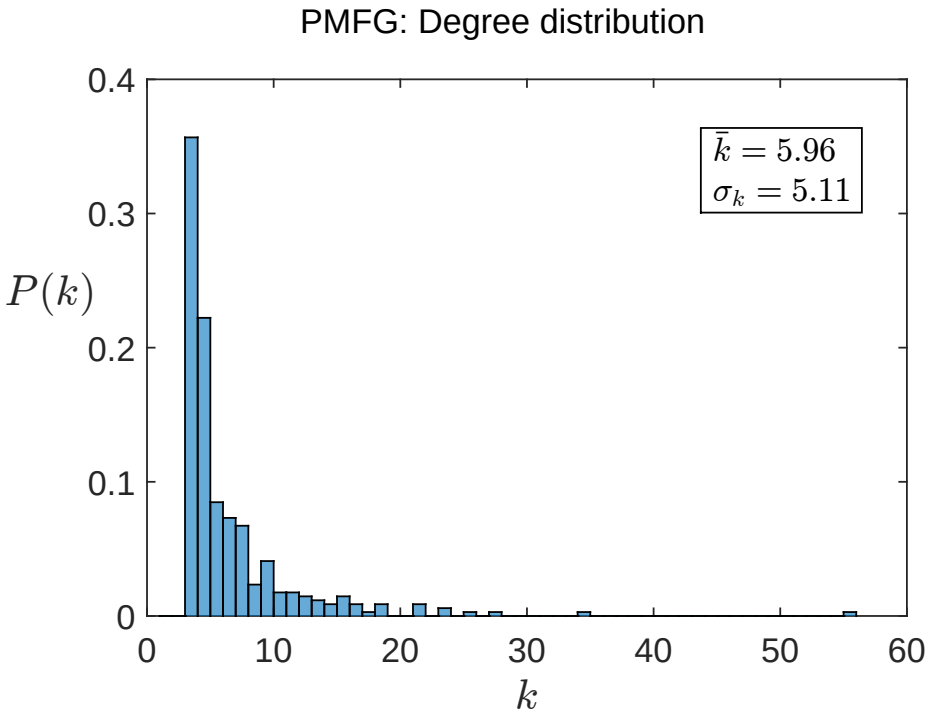

Fig. 3.11 **Degree distribution for the PMFG**. Histogram of degrees calculated from our PMFG. The distribution is again fat-tailed, with a standard deviation of 5.11 greater than both MST and AG ones. The maximum degree is 55.

neighbour degree $k_{nn}(k)$ is an overall decreasing function of k as shown in Fig. 3.12: the PMFG is therefore a disassortative network.

3.3 Insights from Network-filtering: a brief review

Correlation-based networks have provided valuable insights for risk monitoring and portfolio management since the first work by Mantegna [53]. As we have discussed previously, it has been observed that the structure of such networks significantly mirrors the industrial sectors classifications, conveying at the same time important independent information [53, 109]. Such network structure, along with its economic information, is remarkably robust against changes in the sampling frequency of returns time horizon [68] (as long as the market mode is removed from the dependence structure: otherwise

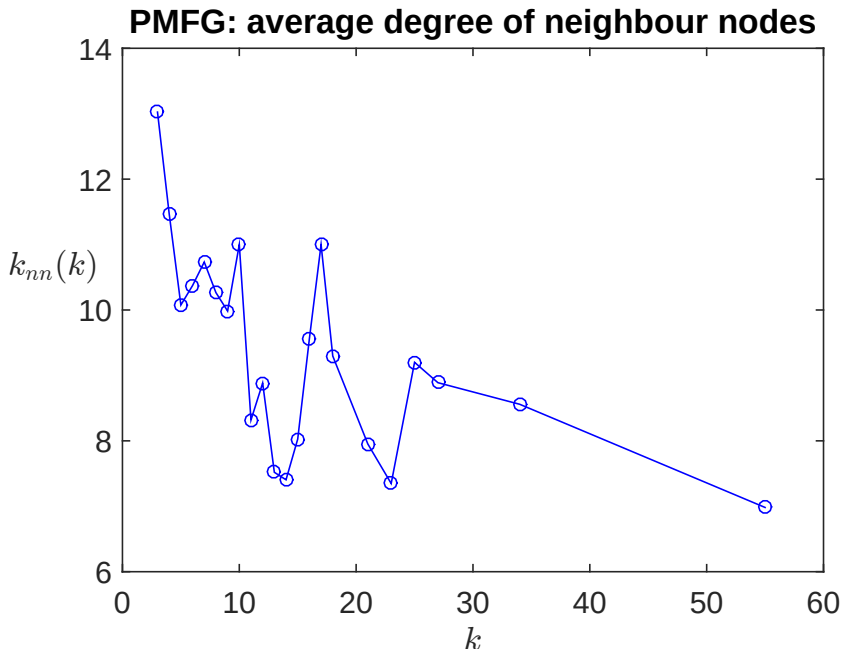

Fig. 3.12 **Average neighbour degree as a function of node's degree in the PMFG**. The average neighbour degree $k_{nn}(k)$ from our PMFG, as a function of node's degree k. As well as the MST case, the overall trend is decreasing, which indicates that the PMFG is a disassortative network: high degree nodes tend to be connected with low degree nodes, and vice-versa.

the structure changes deeply [182, 67]). This has been interpreted as a suggestion "that correlations on short time scales might be used as a proxy for correlations on longer time horizons" [68].

Portfolio optimization methods are improved sensibly by network filtering techniques. For instance, it has been shown in [79] that Markowitz optimization carried out on network-filtered correlation matrices outperforms Markowitz on unfiltered ones. In [81] it has been reported that the peripheral position of nodes in PMFGs can be a criterion to select a well-diversified portfolio. This finding is consistent with what found for the MST in [80], namely that the stocks selected by Markowitz method tend to be the "leaves" of the tree. Eccentricity (a measure of nodes peripherality) has been found to be negatively correlated with asset average return [78], suggesting a connection with the beta factor in CAPM [122]. Correlation-based networks obtained from real data

have also been found to be incompatible with some widespread models for asset returns [34].

Network filtered correlations carry both local and global information in their structure and the analysis of their temporal evolution allows to better understand financial market evolution. For instance, in [70] it has been observed that stocks belonging to the same industrial sector tend to have similar values of centrality in the network topology and that this differentiation is quite persistent over time. In particular, it was observed that Finance, Basic Materials and Capital Goods industrial sectors (Forbes classification) tend to be the located mostly in the central region of the network whereas Energy, Utilities and Health Care are located more in the peripheral region. The preeminent role of the Financial sector is even stronger when correlation networks based on Partial Correlation [183] are analysed [71]; these networks also highlight how the structure of influences among stocks is more complex than the industrial classification, with stocks being affected by several different industrial sectors [72].

Consistently with the cited non-stationarity of financial correlation [38], a certain degree of non-stationarity has been observed on correlation networks too. For instance, the Financial sector appears to loose centrality over the first decade of 2000's [73]. In [75] the authors found both a slow and a fast dynamics in correlation networks topology: while the slow dynamics shows persistence over periods of at least 5 years, the time scale of the fast dynamics is of order of few months and it is linked to special exogenous and endogenous events like financial crises. For instance, in [76] it has been shown that sharp structural changes occurred in the graph topology during the 1987 Black Monday. Similar phenomena have been observed for correlations on Foreign Exchange (FX) data [77]. In [74] it has been demonstrated that structural changes on FX correlation data display different features depending on the type of event affecting the market: news that concern economic matters can trigger a prompt destabilising reaction, whereas when the news consequences for markets are less clear there are periods of "collective discovery" where dynamics appears to gradually synchronise

[74]. In [78] periods of negative returns in equity data are observed to be anticipated systematically by topological modifications of MST - although too many false positive prevent this result being used as a proper forecasting tool.

3.4 Clustering: a complementary perspective on the dependence structure

It turns out there is a deep relation between some correlation-based networks and a class of unsupervised learning techniques, namely the hierarchical clustering methods [15]. In this section we define these methods and elaborate on this connections. In particular we introduce the Directed Bubble Hierarchical Tree algorithm [66], that will play a relevant role in the analyses in the next chapters.

Following notation of Section 3.2, let us call $\{D_{ij}\}$ the $N \times N$ distance matrix (defining pairwise distance among N objects, assets in our analysis) and $\{S_{ij}\}$ the corresponding similarity matrix. $\{D_{ij}\}$ and $\{S_{ij}\}$ are related through the Eq. 3.1 in our case. A clustering method is a technique which, given $\{D_{ij}\}$, groups the N objects in N_{cl} classes ("clusters") so that objects in the same cluster exhibit high similarity among them [62]. The precise criteria according to which the objects are grouped differentiate the clustering methods [62]. The total number of clusters N_{cl} can be either an input or an output of the algorithm, depending on the clustering method. The set of N_{cl} clusters is called "clustering". In the context of Machine Learning, clustering methods belong to the broader class of "unsupervised learning" techniques, because they do not rely on training sets to group new observations [15].

3.4.1 k-medoids

An example of popular clustering technique is the k-medoids method [184]. Its underlying algorithm is the so-called Partitioning Around Medoids (PAM), which is related to that of k-means [185]. In order to identify N_{cl} clusters, PAM works as follows:

1. select randomly N_{cl} "medoids" among the N elements;
2. assign each of the N element to the closest medoid, according to the distance matrix $\{D_{ij}\}$;
3. for each medoid, replace the medoid with each point assigned to it and calculate the cost of each configuration. The cost is defined as the sum of all the distances;
4. choose the configuration with the lowest cost;
5. repeat 2)-4) until no change occurs.

k-medoids can work with any distance matrix $\{D_{ij}\}$, unlike k-means which uses only $L-2$ distance.

3.4.2 Hierarchical Clustering Methods

Among the clustering methods, hierarchical clustering methods (HCM) [63] are especially relevant for Finance [64]. The idea behind HCMs is to compute a hierarchy of N clusterings X^{α} ($\alpha = 1,...,N$) of increasing number of clusters: $N_{cl}^{\alpha} = \alpha$, where N_{cl}^{α} is the number of clusters in the clustering X^{α}. The result is a hierarchical organisation of the N objects which can be represented through a tree diagram called "dendrogram" [62], as shown in Fig. 3.13 for a simplified case of 7 objects. The vertical axis in the dendrogram represents a distance: when two clusters are merged together, the corresponding value on the vertical axis corresponds to the distance between the two clusters. From this hierarchy of clusterings a single clustering can be obtained by choosing a number of clusters N_{cl} (that is therefore a free parameter) and cutting the dendrogram at the appropriate level of distance: this is shown in Fig. 3.13, where two cuts at different levels generate two clusterings with $N_{cl} = 2$ and $N_{cl} = 3$ respectively. This approach is deeply different from non-hierarchical clustering methods such as k-medoids, where no hierarchical relation exists among clusterings generated with different N_{cl}, and therefore no dendrogram is created.

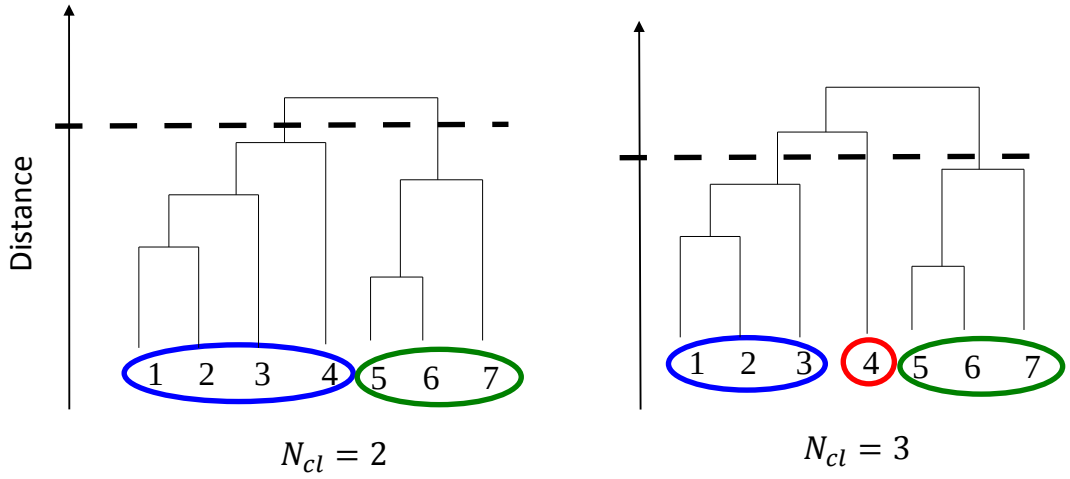

Fig. 3.13 **Selection of two clusterings from a dendrogram of 7 objects**. Outline of the procedure to obtain a clustering from a hierarchical tree (dendrogram) generated by a hierarchical clustering technique. By cutting the tree at the appropriate level of distance we obtain a clustering made of 2 (left graph) and 3 (right graph) clusters respectively.

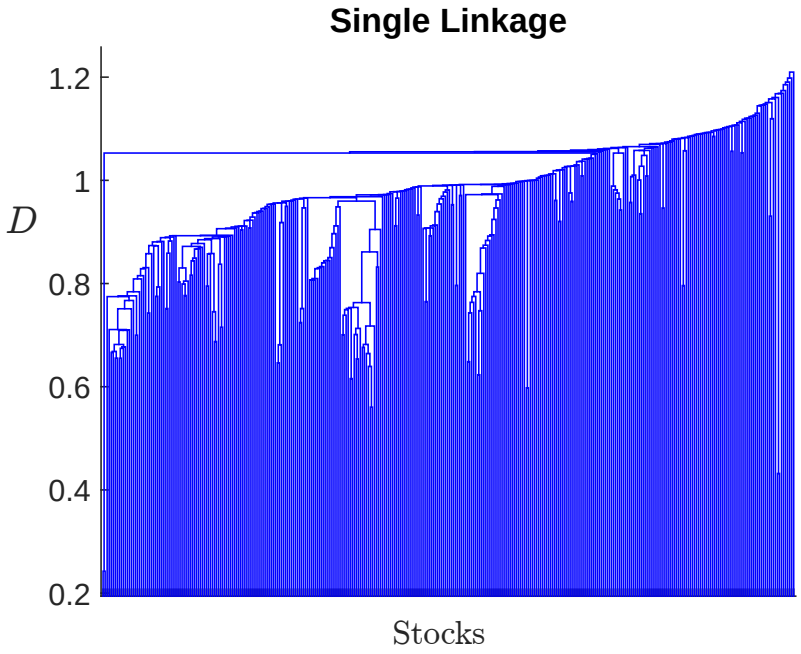

Fig. 3.14 **Dendrogram generated by the Single Linkage method**. Hierarchical tree obtained by performing the SL clustering on the correlation matrix $\{\rho_{ij}\}$ of the data set.

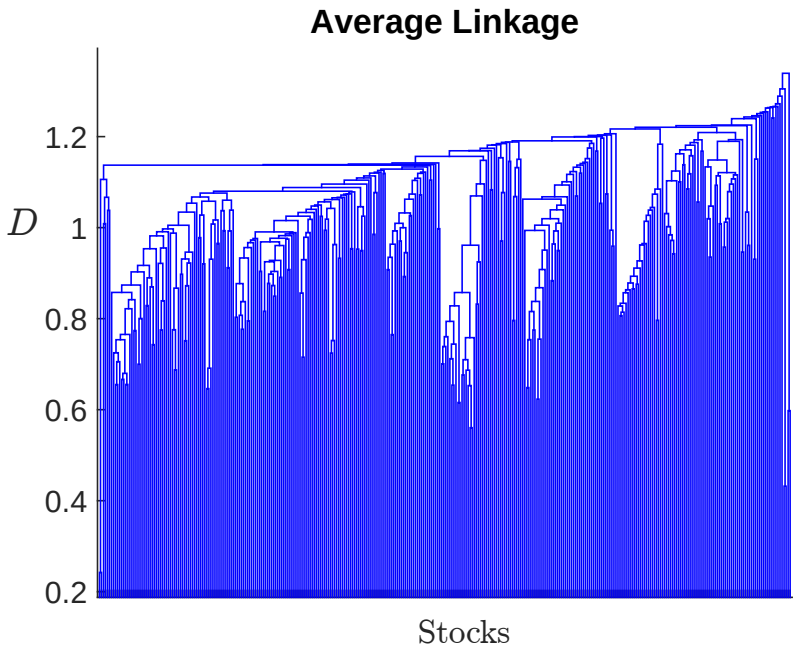

Fig. 3.15 **Dendrogram generated by the Average Linkage method**. Hierarchical tree obtained by performing the AL clustering on the correlation matrix $\{\rho_{ij}\}$ of the data set.

In the following subsections we describe in more details some of the main HCMs used in Finance [64].

3.4.3 Linkage methods

Linkage methods are a family of HCMs where the distance among clusters is defined in terms of the original distance matrix $\{D_{ij}\}$. Given $\{D_{ij}\}$, Linkage methods define first a set of N clusters, each one made of one object only; the distance between two clusters is simply the distance between the two corresponding object in $\{D_{ij}\}$. Then the closest (i.e. least distant) pair of clusters is merged into a new cluster. The latter step is repeated until only one cluster remains, made of all the objects. The distance among two generic clusters A and B is at each step defined and updated according to a different formula depending on the specific Linkage method. They are also called

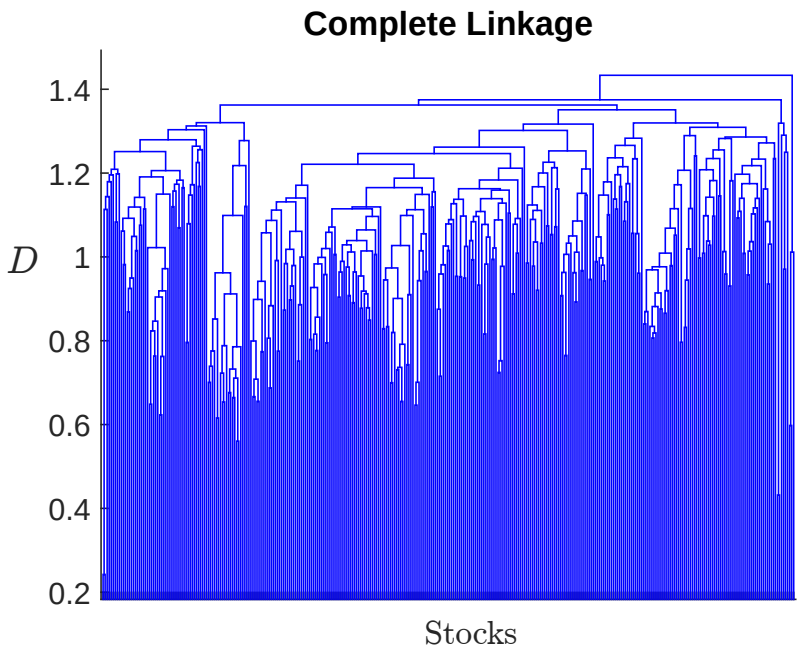

Fig. 3.16 **Dendrogram generated by the Complete Linkage method**. Hierarchical tree obtained by performing the SL clustering on the correlation matrix $\{\rho_{ij}\}$ of the data set.

"agglomerative" clustering methods, since they begin with a partition of N clusters and then proceed merging them.

In **Single Linkage** (SL) [62, 64] the distance between clusters A and B is at each step defined as:

$$d_{AB} = \min_{a \in A, b \in B} D_{ab} \ . \tag{3.5}$$

The Linkage procedure so defined resembles the algorithm for constructing a MST from $\{D\}_{ij}$, as described in the Section 3.2. In fact it can be shown [64] that the MST algorithm is basically the SL procedure carried out until the graph is completely connected. There is therefore a strict relation between the two tools: the MST can be seen as a network representation of the hierarchy generated by the SL, although it retains some information that the SL discards [64]. This hierarchy defines a new metrics

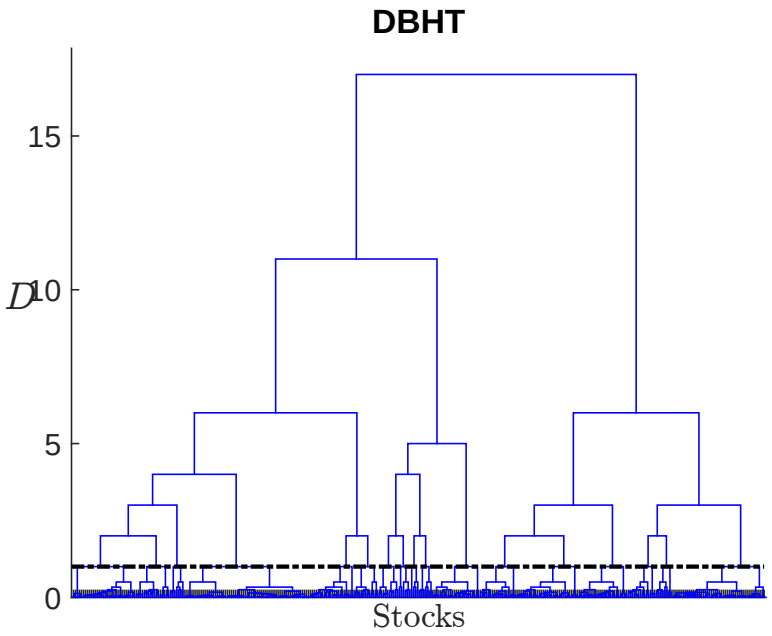

Fig. 3.17 **Dendrogram generated by the Direct Bubble Hierarchical Tree method**. Hierarchical tree obtained by performing the DBHT clustering on the correlation matrix $\{\rho_{ij}\}$ of the data set. The horizontal dotted line identifies the natural clustering extracted from the method, which in this case is made of 17 clusters.

among the N nodes [53], that is an ultrametric distance [186] where the condition in Eq. 3.4 is replaced by the stronger condition:

$$D^{<}_{ij} \leq \max(D^{<}_{ik}, D^{<}_{kj}) \quad . \tag{3.6}$$

The new distance $D^{<}_{ij}$ can be calculated from the MST, as it equals the maximum (metric) distance D detected by moving from i to j through a shortest path. Hence the MST maps the original metric space in a new, ultrametric space, to which a unique hierarchical organization of the nodes corresponds. Moreover, the corresponding ultrametric space provides information on the risk factors affecting each asset [53].

Different rules other than Eq. 3.5 can be used to update the distance between clusters. In the **Average Linkage** (AL) [62, 64] algorithm Eq. 3.5 is replaced by:

$$d_{AB} = mean_{a\in A, b\in B} D_{ab} \quad . \tag{3.7}$$

Also AL has been shown to be associated to a filtering network, namely a slightly different version of spanning tree [140], called Average Linkage Minimum Spanning Tree. Finally the **Complete Linkage** (CL) [62, 64] is a third variant of Linkage, where Eq. 3.5 is replaced by:

$$d_{AB} = \max_{a\in A, b\in B} D_{ab} \quad . \tag{3.8}$$

Different Linkage methods can yield very dissimilar dendrograms. In Figs. 3.14 - 3.16 we show the dendrograms which we have obtained by performing respectively the SL, the AL and CL methods to the equity data set. We find that the tree structure of SL in Fig. 3.14 is characterised by the emergence of a very large cluster, that is gradually joined by single stocks (or very small clusters): visually, the process can be observed as a gradual build-up of stocks from the left to the right of the graph. The CL dendrogram in Fig. 3.16 is instead characterised by a more homogeneous clusters size distribution for all values of distance, as evident from the more symmetric tree structure. Finally, AL dendrogram in Fig. 3.15 seems to be an intermediate case between SL and CL. We will discuss in more detail the differences between Linkage methods on financial data in Chapter 4.

3.4.4 Directed Bubble Hierarchical Tree

As we have seen, the MST is deeply connected to a clustering method, namely the Single Linkage algorithm. Since the PMFG is a generalization of the MST, it could be raised the question whether to the PMFG corresponds a clustering method that exploits this higher amount of information. In [66] it has been shown that this is the case: the PMFG topology, due to its property of being made of three-cliques, defines a hierarchy over the set of nodes [83, 66] that can be revealed and used to group them

in communities. The clustering algorithm that exploits this property is called Directed Bubble Hierarchical Tree (DBHT) [66]. In particular the DBHT exploits the distinction between separating and non-separating three-cliques to identify a clustering partition of all the nodes in the PMFG [66]. A complete dendrogram is then obtained both inter-clusters and intra-clusters by following a traditional agglomerative clustering procedure. A more detailed description of DBHT algorithm is given in Appendix A.

Since DBHT exploits the topology of the correlation network, it can be viewed as an example of community detection algorithm in graphs [187]. It is worth noting the difference between Linkage algorithms and DBHT. Linkage algorithms look at the sorted list of distances D_{ij} and then build the dendrogram by gathering subsets of stocks with lowest distances; the community partition is then obtained, as we said, from the dendrogram after choosing the parameter N_{cl} "number of clusters". The DBHT instead reverses this order: first a "natural clustering" is identified by means of topological considerations on the planar graph, with the corresponding N_{cl} that is therefore an output of the method; then from this clustering a dendrogram is constructed both inter-clusters and intra-clusters. The difference involves therefore both the kind of information exploited and the methodological approach.

In Fig. 3.17 we show the dendrogram obtained by performing the DBHT on the data set. The horizontal dotted line identifies the natural clustering provided by the method, made of 17 clusters in this case. In terms of clusters size the DBHT dendrogram is similar to the CL one, with quite an homogeneous size distribution and symmetric dendrogram. We will discuss the DBHT hierarchical structure in more detail in Chapter 4.

3.5 Summary

In this chapter we have introduced the concept of correlation-based filtered networks. Specifically we have explained the principles underlying the construction of MST, AG,

EG and PMFG from correlation matrices, and highlighted the differences among these approaches. By applying network filtering to the equity data set we have demonstrated that this tool can extract meaningful information from the correlation matrix; the analysis of topology has revealed a complex structure, characterised by fat tails in the degree distribution and degree-degree correlation.

We have then showed how the use of correlation-based networks has allowed to address many questions of interest in Quantitative Finance, such as portfolio optimization, risk diversification, non-stationarity, the relation between news and prices and the dynamic of financial crises. Their versatility lies in the power of Network Theory, that is explicitly conceived to address multi-dimensional problems and non-trivial interactions.

Notably, some of these network representations turn out to be closely related to hierarchical clustering methods, such as Linkage and the Directed Bubble Hierarchical Tree techniques. In a way, correlation-based networks can be viewed as visual representations of the cluster communities, although they also provide independent information. In the next chapter we elaborate on this connection and we analyse in depth the clustering structure associated to network filtering on financial data.

Chapter 4

Relation between financial market data and real economy

In this chapter we apply for the first time the Directed Bubble Hierarchical Tree (DBHT) method to financial data. Through this clustering method we investigate how the dependence structure of stocks relates to the underlying economic activity, as represented by the stocks industrial sector membership. Furthermore, we analyse how such relation depends on time and is affected by turbulent market periods. We investigate these questions by comparing the DBHT with other clustering methods as well. Part of the results and analyses presented in this chapter has been published in the paper "Relation between Financial Market Structure and the Real Economy: Comparison between Clustering Method" in 2015 [109].

4.1 Introduction

It is a long known fact that the dependence structure partially mirrors the industrial sector classification [118]. This fact supports the intuitive argument that returns of stocks in the same industrial sector are affected mainly by the same flows of information and economic environment. However, to the best of our knowledge this relation has been explored only qualitatively so far [154]. An exception is [68], where however only

one clustering method is analysed. In [106] different clustering and spectral methods are compared quantitatively in terms of amount of filtered information: yet this comparison was performed without looking at the industrial sector classification, by assuming a multivariate Gaussian distribution for the stocks returns [64].

In this chapter we describe a set of analyses that aim to quantify the relation between correlation and industrial classification without any assumption on returns distribution. This is a relevant improvement since multivariate Gaussian models are known to be inaccurate to describe stocks returns [127, 3]. To this end, we perform a hierarchical cluster analysis on the empirical correlation matrix, and investigate the relation between the dendrogram structure and the industrial membership of the stocks. Our analyses are also dynamical and include comparisons among different clustering methods.

The original contributions of this chapter are the following:

- We apply for the first time the DBHT method to financial data and we highlight its advantages over other clustering methods.
- We quantify and compare the degree of economic information extracted from five clustering methods by means of the Adjusted Rand Index and the Hypergeometric hypothesis test, revealing that different methods extract quite different information.
- We perform a dynamical analysis of the clusterings structure. We find that the choice of the clustering method affects the sensitivity of the clustering structure to financial crises.
- We find that the market mode detrending affects some clustering methods more than others. By dynamically comparing detrended and non-detrended DBHT clustering we find evidence that the market mode influence has steadily increased over the period 1997-2012.

This chapter is organised as follows. In Section 4.2 we analyse the clustering structure related to the correlation matrix in a static setting, with a single time window

covering the whole 15 years period. In particular in Subsection 4.2.1 we analyse the clustering composition in terms of industrial sectors for each clustering method, whereas in Subsection 4.2.2 we investigate the clusters size distribution for each method. The amount of economic information is then quantified in Subsections 4.2.3 and 4.2.4. In Section 4.3 we perform instead analyses using a dynamical approach, with moving time windows. In particular the evolution of the economic information in the dependence structure is analysed in Subsection 4.3.1.

4.2 Structure and economic information of the correlation clustering

In this section we investigate the clustering structure of our stocks data through a variety of tools. In particular we focus on the economic information contained in the dendrogram, which we measure by studying how stocks from the same industry are related in the hierarchical structure. The analyses will be presented first in a static way, that is by calculating correlation on the whole period 1997-2012; then we will turn to a moving window set-up that allows to explore the dynamic evolution.

The clustering methods we use are the five introduced in Chapter 3: the three Linkage methods (SL, AL and CL) [62], the DBHT [66] and the k-medoids [184]. All of them provide a clustering for each choice of the number of clusters N_{cl}; however, only the Linkage and DBHT methods generate a dendrogram.

4.2.1 Clusters composition

We begin by looking at the clusterings composition in terms of ICB supersectors and industries. The DBHT method yields 17 clusters. The associated dendrogram is shown in Fig. 3.17. We can characterize each cluster in terms of its industrial sector composition. In Fig. 4.1 to each cluster is associated a bar, whose height represents the number of stocks in the cluster. Each bar is made of different colors, showing the

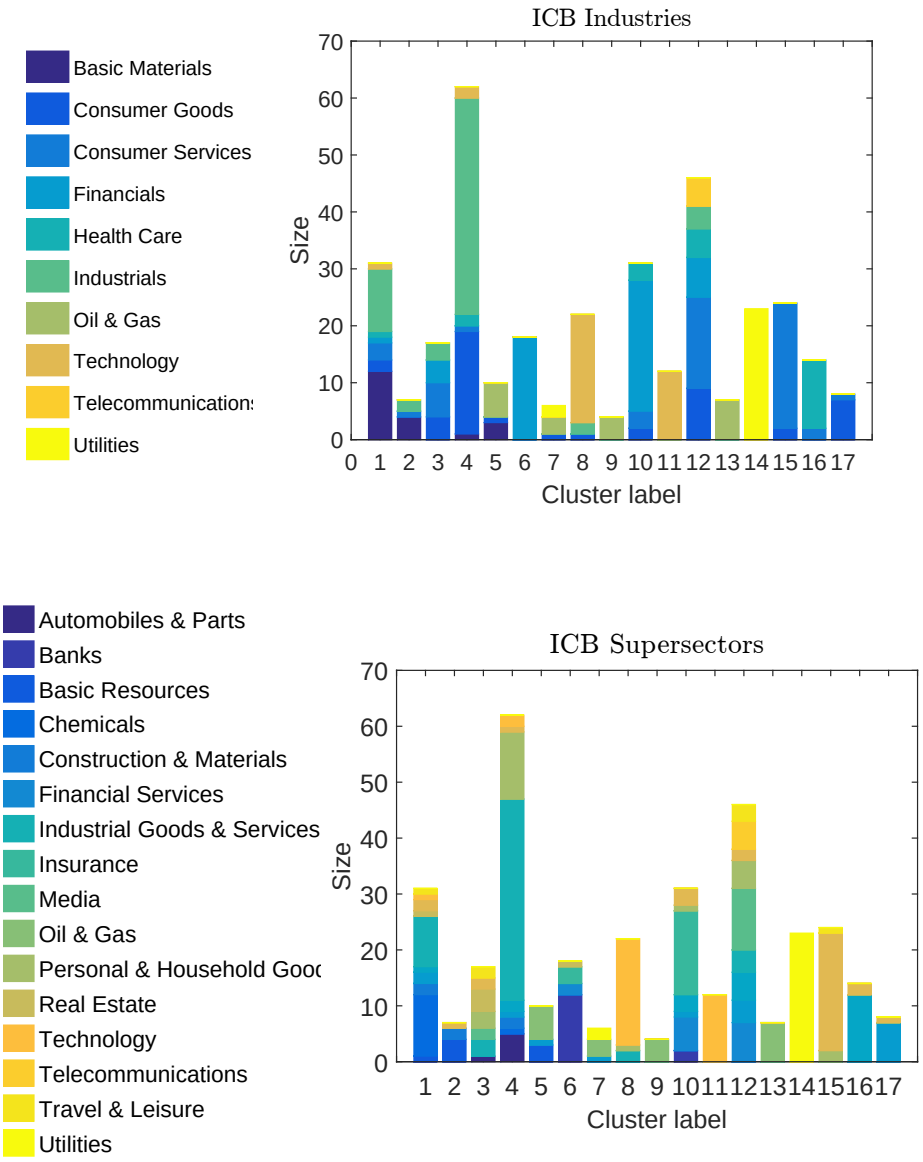

Fig. 4.1 **DBHT clustering composition**. Upper graph: DBHT clustering composition in terms of ICB industries. Bottom graph: DBHT clustering composition in terms of ICB supersectors.

composition of each cluster in terms of ICB industries (upper graph) and supersectors (lower graph). Cluster 4, the largest, is made of 62 stocks, accounting for about the 18% of the total number of stocks; cluster 9, the smallest, contains 4 stocks. The average size of clusters is 20.1 stocks.

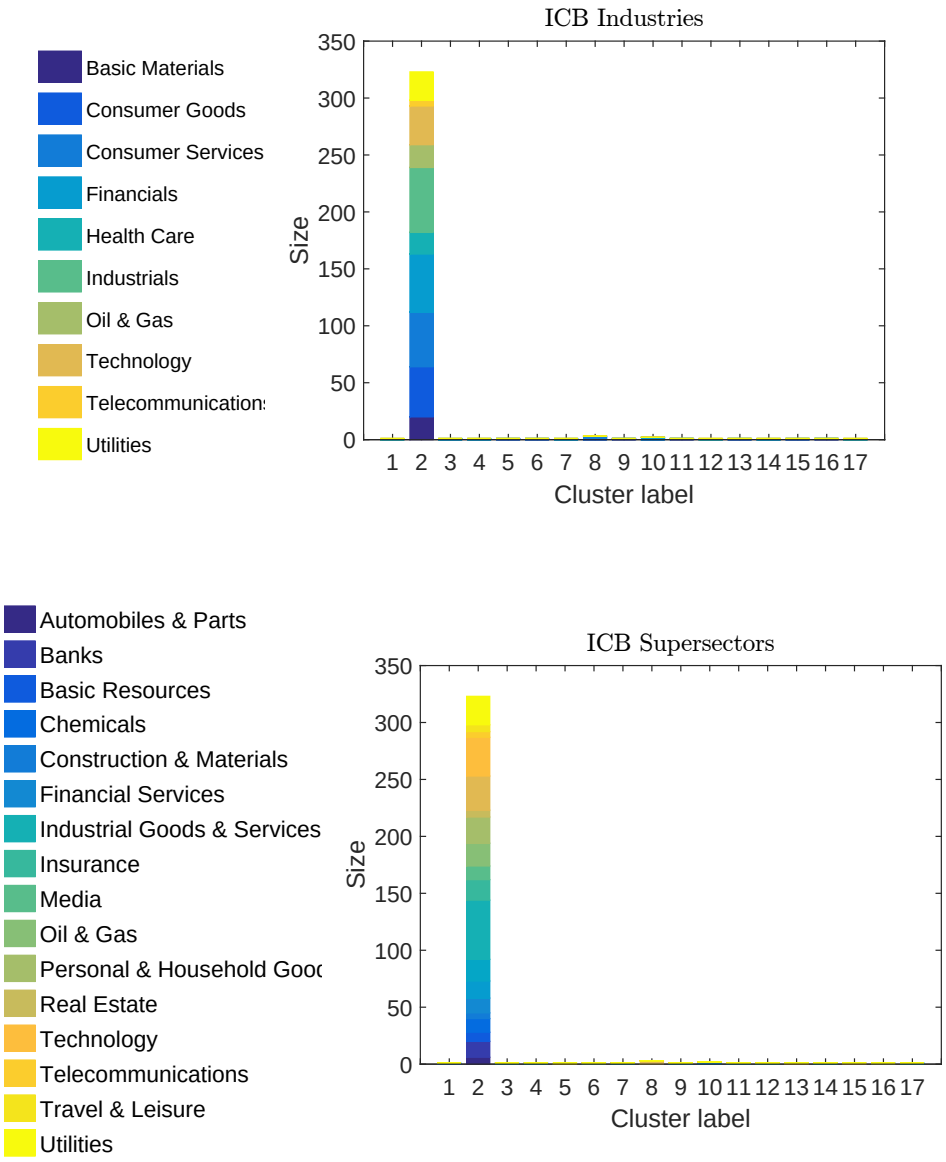

Fig. 4.2 **Single Linkage clustering composition**. Upper graph: SL clustering composition in terms of ICB industries. Bottom graph: SL clustering composition in terms of ICB supersectors.

As we can see, four clusters show a composition of stocks belonging to only one ICB supersector: cluster 9 and 13 (Oil & Gas), 11 (Technology) and 14 (Utilities). Similar cases are cluster 8, made of Technology stocks for more than 86%, cluster 15, within which 91% of stocks are from Retail, cluster 16 (75% of stocks from Health Care) and cluster 17 (87.5% of stocks from Food & Beverage). Moreover there are

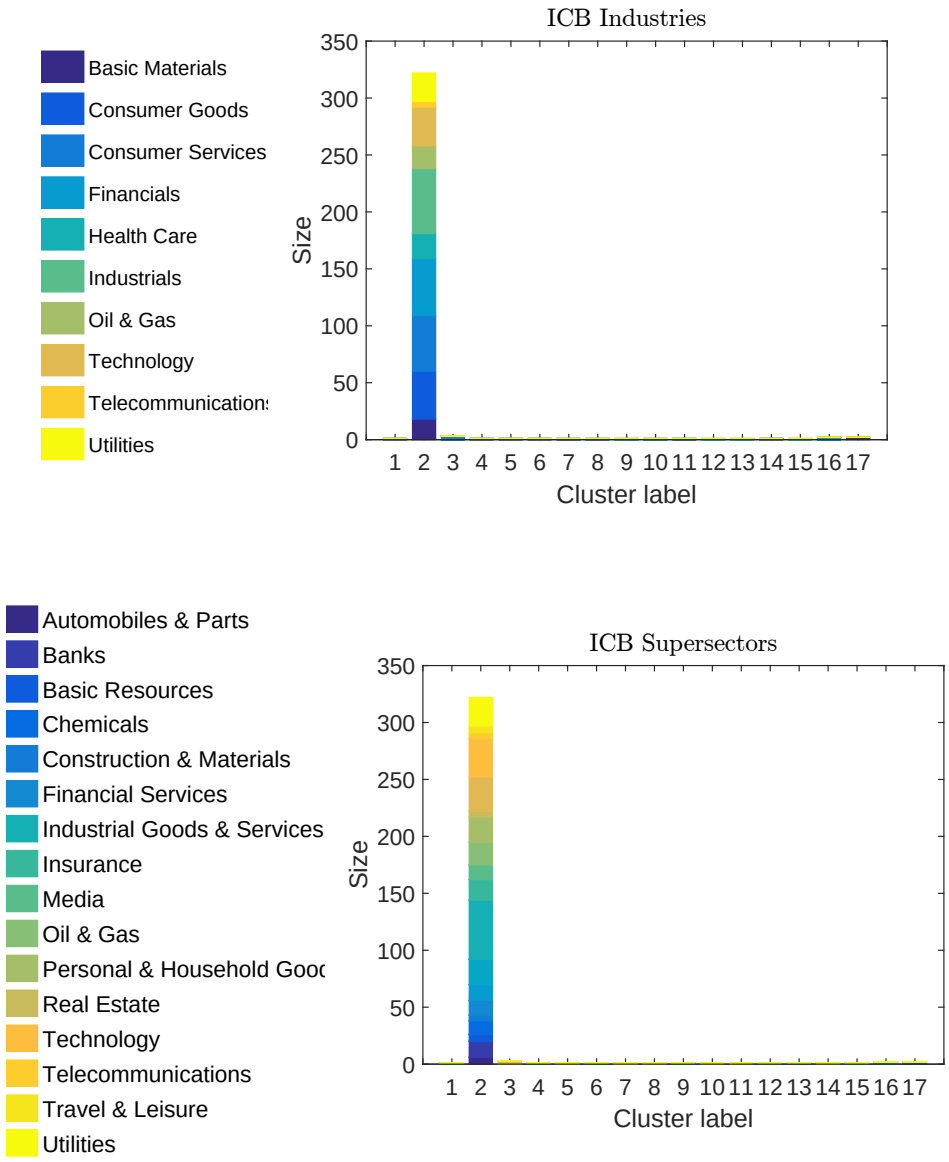

Fig. 4.3 **Average Linkage clustering composition**. Upper graph: AL clustering composition in terms of ICB industries. Bottom graph: AL clustering composition in terms of ICB supersectors.

clusters that, although showing a mixed composition, are composed by supersectors strictly related: the number 6 is made of Banks, Financial Services and Insurance, all supersectors that the ICB gathers in the same industry (Financial) at the superior hierarchical step.

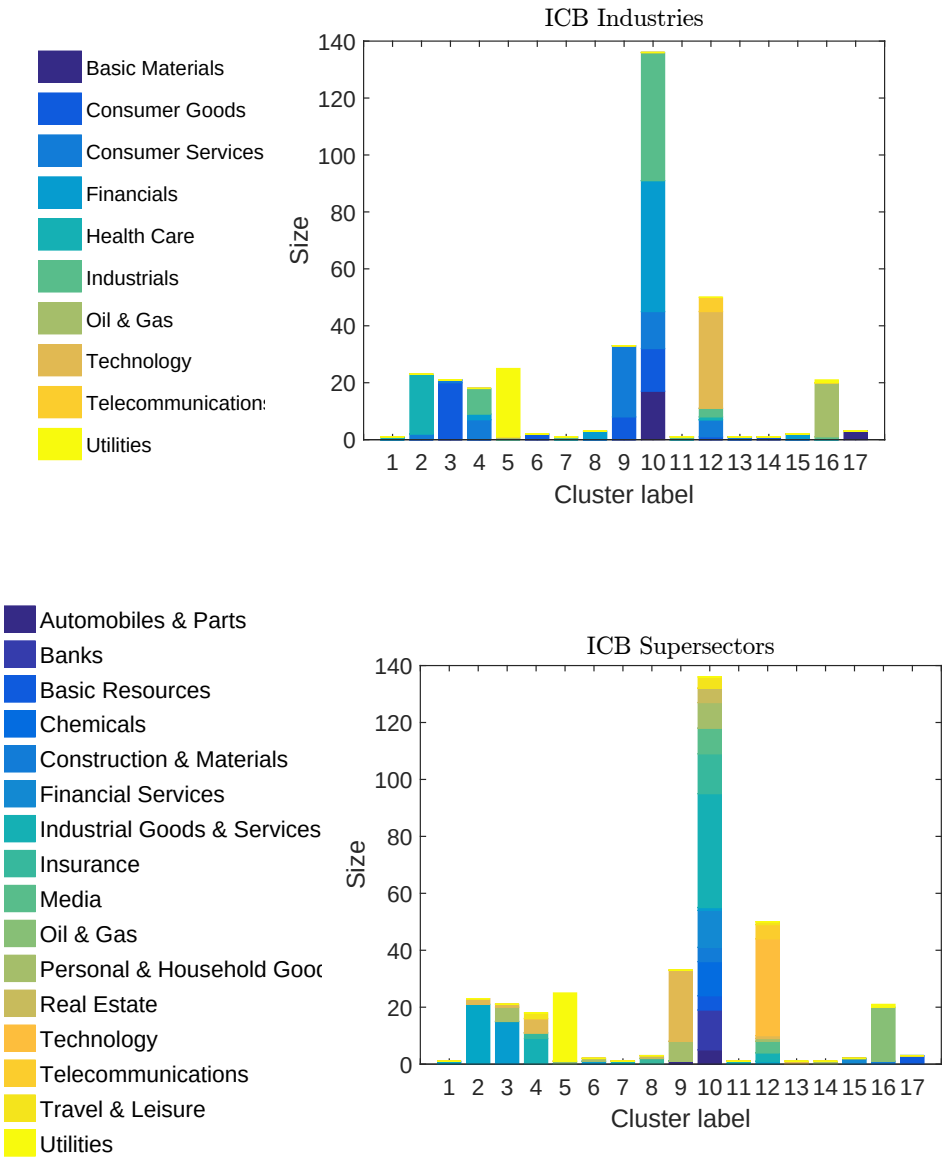

Fig. 4.4 **Complete Linkage clustering composition**. Upper graph: CL clustering composition in terms of ICB industries. Bottom graph: CL clustering composition in terms of ICB supersectors.

There are clusters that do not show an overexpression for a particular supersector or industry: this fact points out that the clustering is after all providing an information that cannot be reduced only to the industrial classification. In particular clusters 1, 3 and 12 have a heterogeneous composition, covering almost all the 19 supersectors and with no sector dominating the others. The cluster 4 is an intermediate case, since even though it

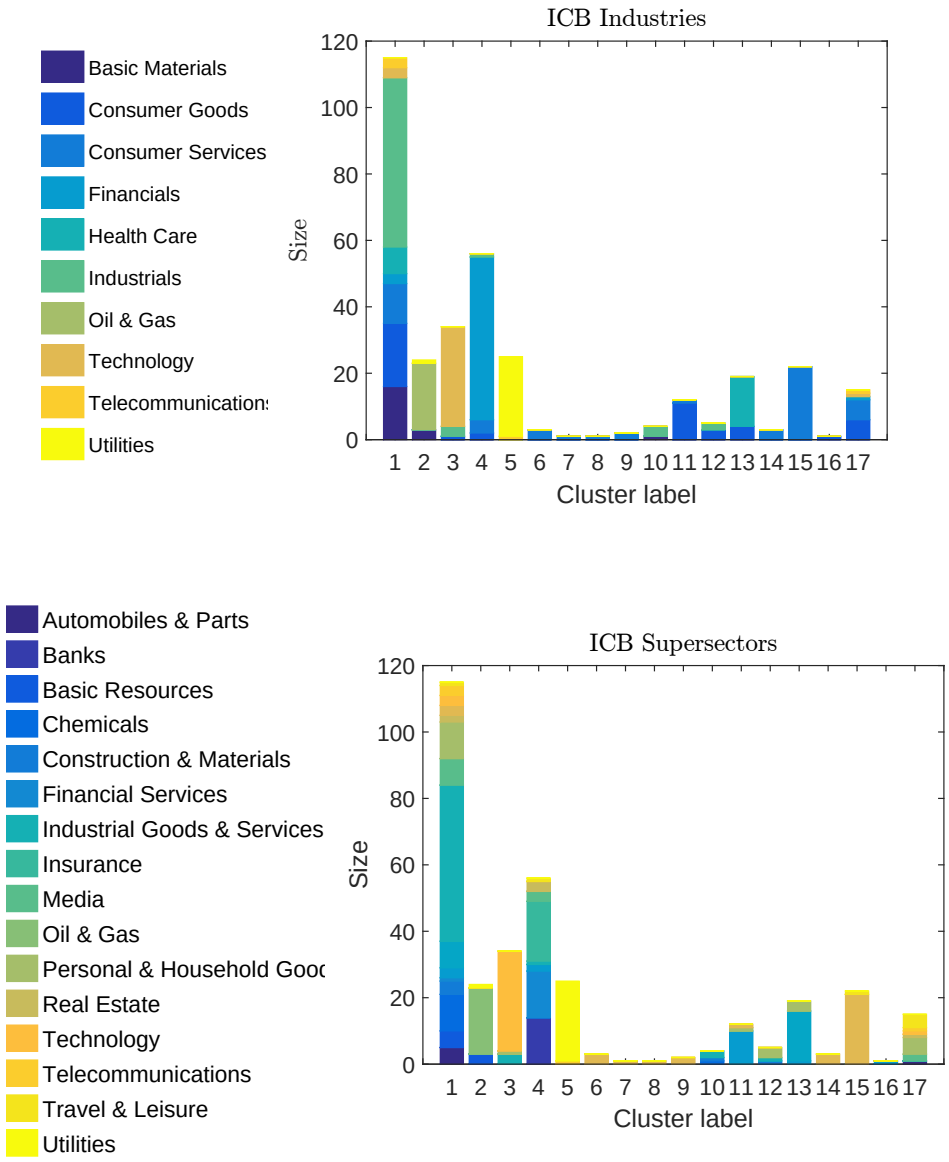

Fig. 4.5 **k-medoids clustering composition**. Upper graph: k-medoids clustering composition in terms of ICB industries. Bottom graph: k-medoids clustering composition in terms of ICB supersectors. The number of clusters is chosen to be 17, equal to the number of DBHT clusters.

overexpresses the Industrial Goods & Services (75%), it contains stocks belonging to 9 different supersectors and 6 industries. The largest clusters (4, 12, 1 and 10) are all among these types of "mixed" clusters.

Let us now focus on the Linkage methods. The number of clusters for these methods has been chosen equal to 17, in order to compare the results with those for DBHT. We show in Fig. 4.2, 4.3 and 4.4 the obtained clusters compositions. First of all we can observe that SL and AL display a strong heterogeneity in the size of clusters: they have two huge clusters of 323 and 322 stocks respectively (almost identical, having 318 stocks in common), with the other clusters made of one, two or three stocks. For both the algorithms this giant cluster contains stocks of all ICB sectors.

For what concerns the CL, the situation is quite different. The giant cluster (cluster number 10) is much reduced in size (136 stocks), with also other three clusters (the number 12, 9 and 5) containing a relevant number of stocks (50, 33 and 25 respectively): the main supersectors that are overexpressed are Technology (cluster 12), Utilities (cluster 5), Retail (cluster 9), Oil & Gas (cluster 16) and Health Care (cluster 2). A very similar structure occurs with the k-medoids in Fig. 4.5, but with the giant cluster further reduced in size. However the DBHT clustering is the one showing the largest degree of homogeneity in size and overexpression of ICB industries and supersectors, at least for this number of clusters.

Clusters composition: detrended log-returns

Let us now describe how the above structures change when clusterings are obtained from the detrended correlation matrix $\rho^R(T)$, introduced in Chapter 2. For the DBHT the number of clusters is now 23. The largest cluster contains 45 stocks (13% of total), the smallest 4. The average size is 14.8. As we can see, some supersectors that were mixed together in the non-detrended case are now overexpressed in distinct clusters: Chemicals (cluster 1), Insurance (cluster 11) and Telecommunications (cluster 19). In terms of Industries, a cluster made entirely of Consumer Goods stocks appears (cluster 21). However, some supersectors that were overexpressed in the non-detrended case now tend to be more spread over different clusters: Utilities, Oil & Gas, the Financial

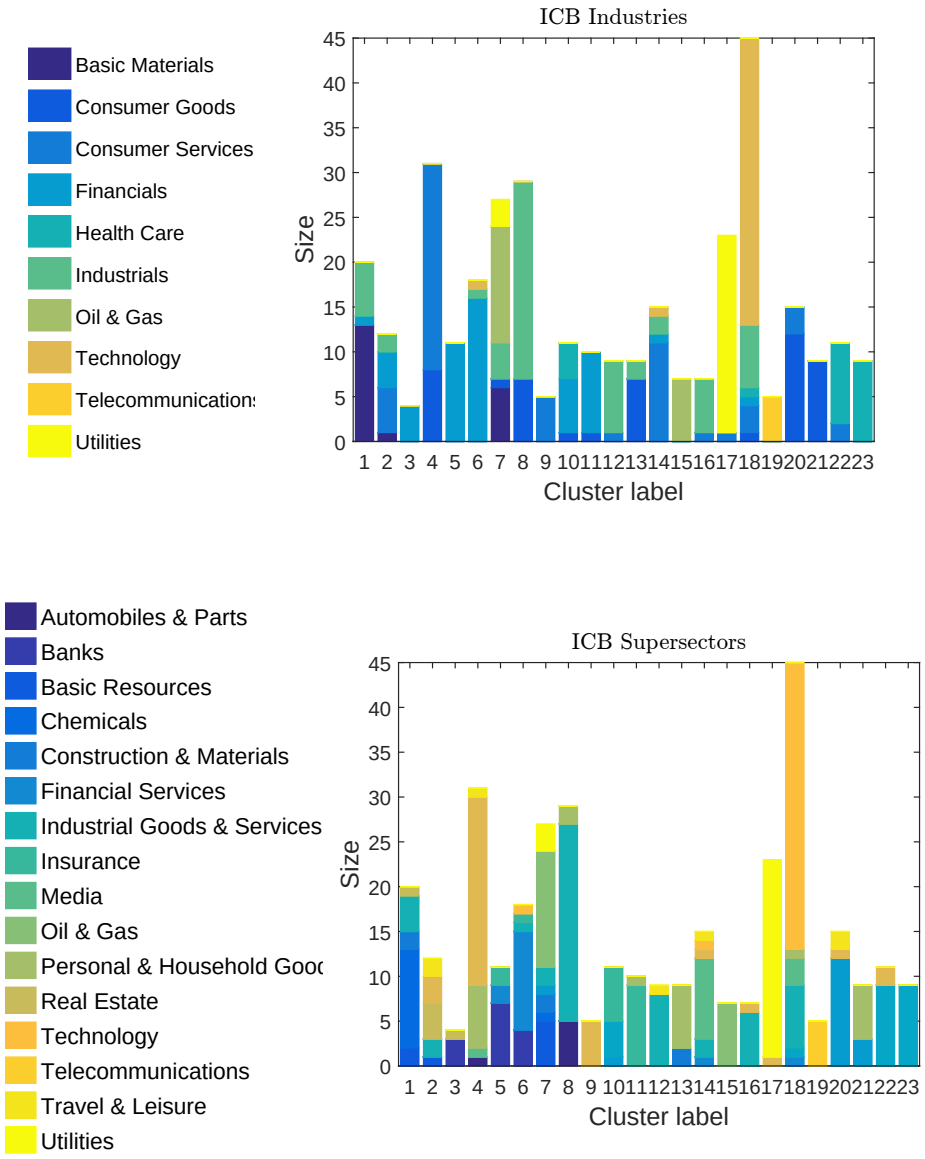

Fig. 4.6 **DBHT clustering composition from detrended log-returns**. Upper graph: DBHT clustering composition in terms of ICB industries. Bottom graph: DBHT clustering composition in terms of ICB supersectors. Both clustering are computed from log-returns detrended of the market mode.

industry. This again points out the difference between the dependence structure and economic based classifications such as ICB.

The clusters composition from Linkage methods are shown in Figs. 4.7 - 4.9. We can observe that for SL there is still a strong heterogeneity in the size of clusters, with

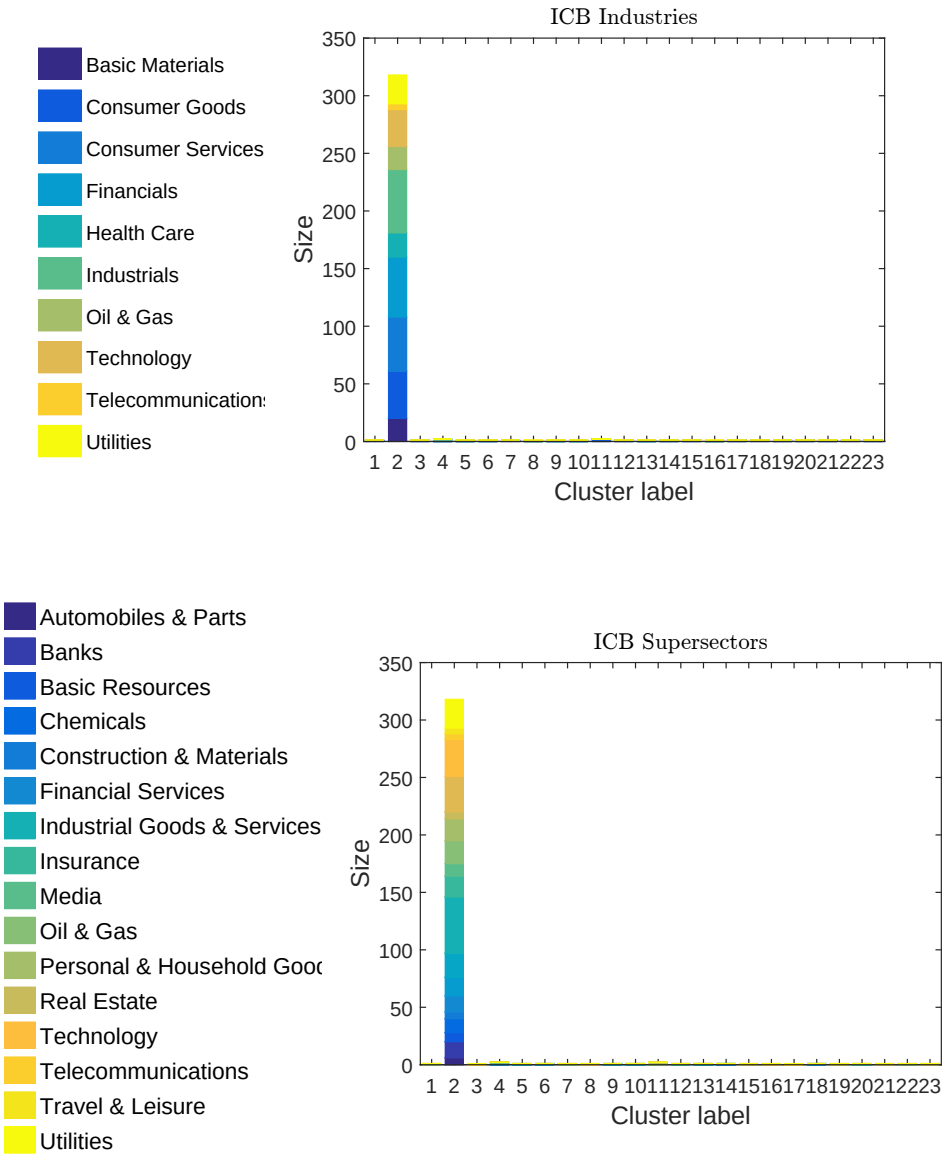

Fig. 4.7 **Single Linkage clustering composition from detrended log-returns**. Upper graph: SL clustering composition in terms of ICB industries. Bottom graph: SL clustering composition in terms of ICB supersectors. Both clustering are computed from log-returns detrended of the market mode. The number of clusters is chosen to be 23, equal to the number of DBHT clusters.

the presence of a giant cluster containing 318 stocks. On the contrary, AL displays now a more structured clustering: the size of the largest cluster shrinks to 58 stocks, and 6 different clusters of medium size (20-40 stocks) appear. Moreover, these clusters show a much higher overexpression of supersectors than in the non-detrended case, such as

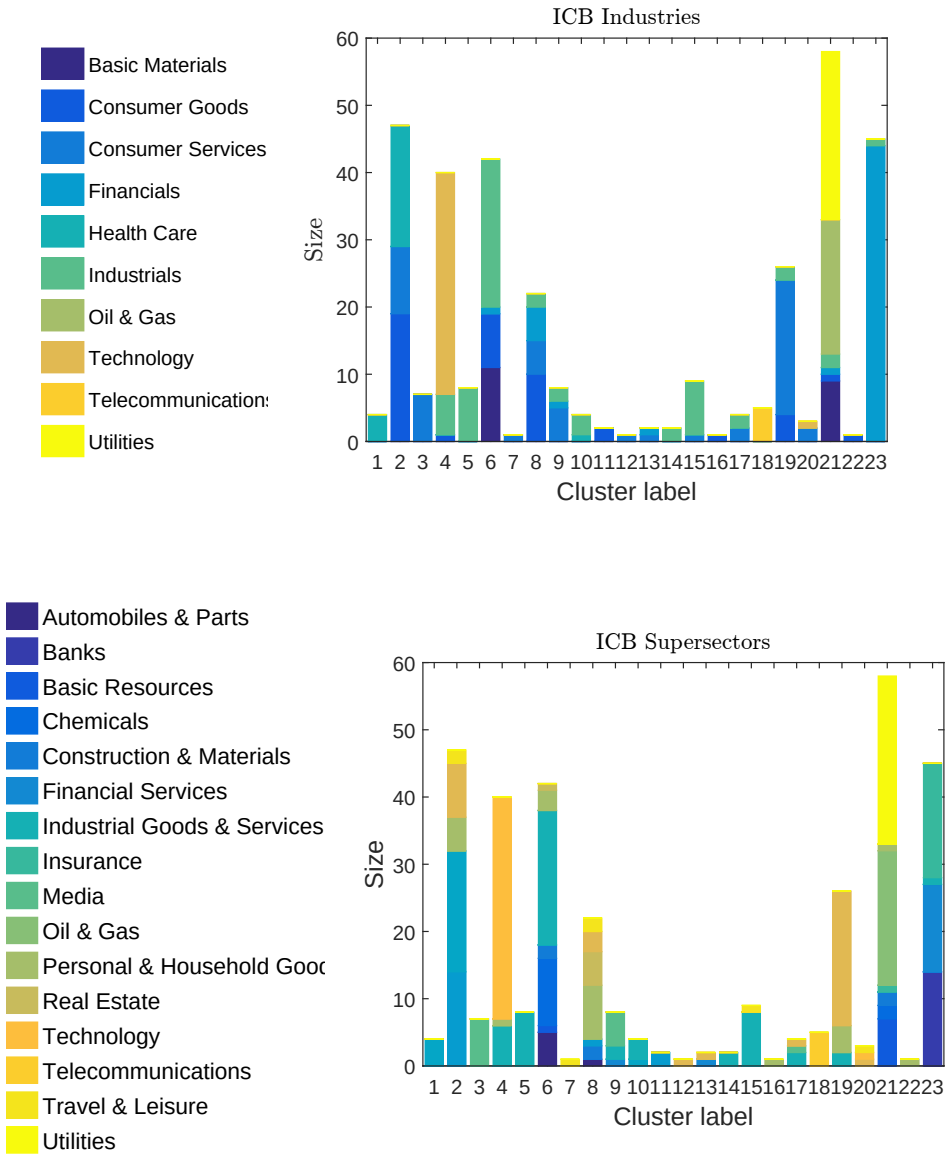

Fig. 4.8 **Average Linkage clustering composition from detrended log-returns**. Upper graph: AL clustering composition in terms of ICB industries. Bottom graph: AL clustering composition in terms of ICB supersectors. Both clustering are computed from log-returns detrended of the market mode. The number of clusters is chosen to be 23, equal to the number of DBHT clusters.

Technology (cluster 4), Industrial Goods & Services (cluster 5 and 15), Media (cluster 3), as well as Finance industry (cluster 23). However there are still 10 clusters whose size is at most 4 stocks. For the CL and the k-medoids the supersectors overexpression is further improved, becoming as rich as the DBHT one. Especially CL shows overexpression of

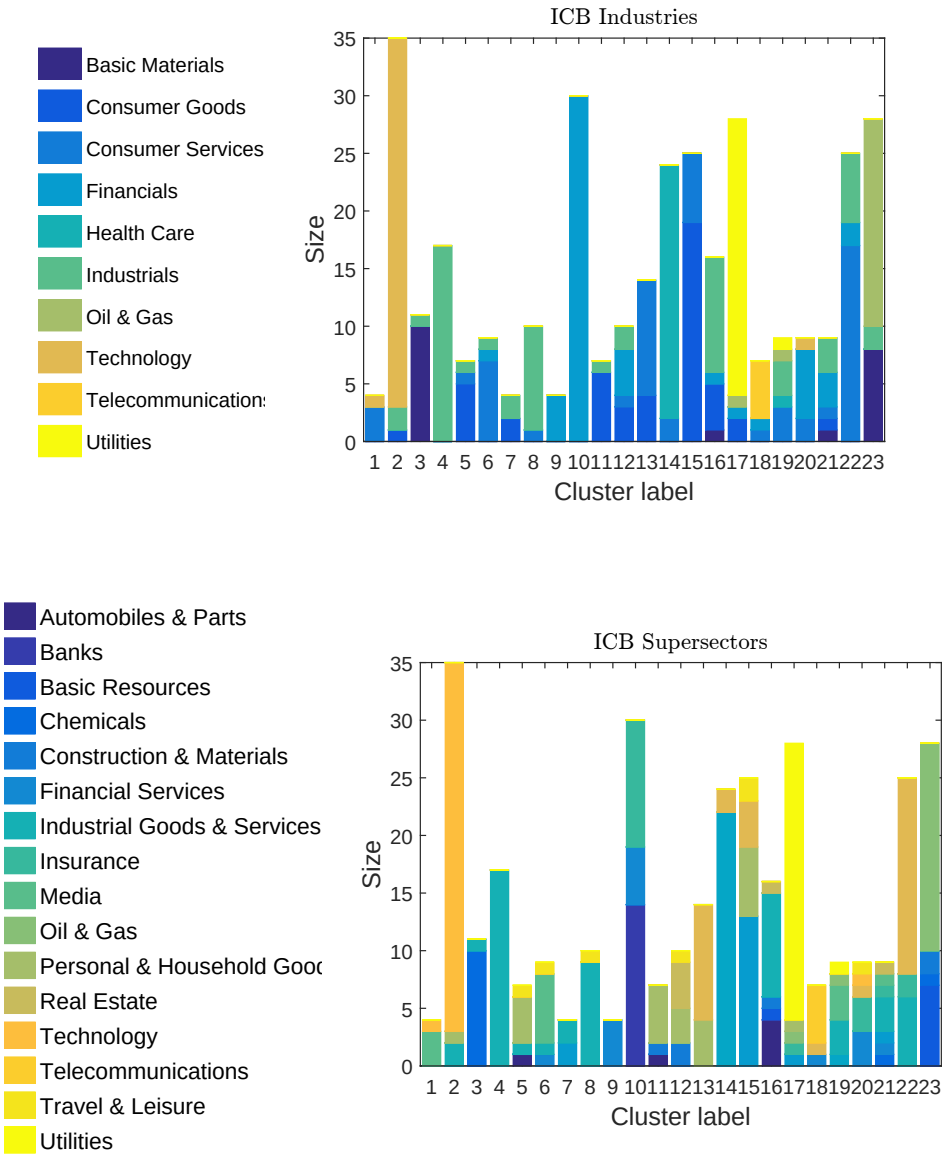

Fig. 4.9 **Complete Linkage clustering composition from detrended log-returns**. Upper graph: CL clustering composition in terms of ICB industries. Bottom graph: CL clustering composition in terms of ICB supersectors. Both clustering are computed from log-returns detrended of the market mode. The number of clusters is chosen to be 23, equal to the number of DBHT clusters.

Technology (cluster 2), Industrial Goods & Services (clusters 4 and 8), Utilities (cluster 17), Oil & Gas (cluster 23), Health Care (cluster 14) and Financial Services (cluster 9). Similar overexpressions are found for the k-medoids case, in Fig. 4.10.

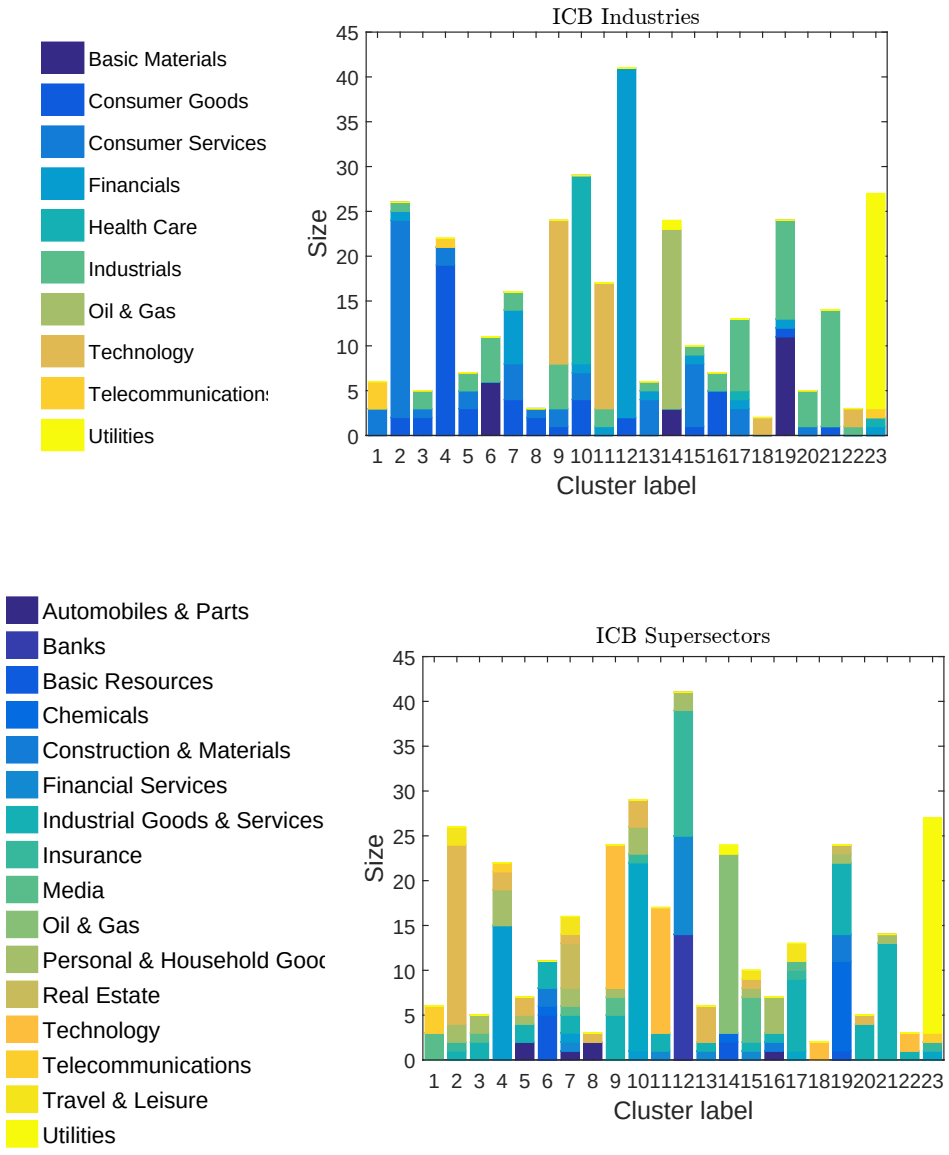

Fig. 4.10 **k-medoids clustering composition from detrended log-returns**. Upper graph: k-medoids clustering composition in terms of ICB industries. Bottom graph: k-medoids clustering composition in terms of ICB supersectors. Both clustering are computed from log-returns detrended of the market mode. The number of clusters is chosen to be 23, equal to the number of DBHT clusters.

Overall, we can conclude that by subtracting the market mode we get a richer, more structured clustering that shows an higher amount of ICB related information. The sensitivity to this change is strongly dependent on the clustering method: in particular SL is not affected, whereas AL shows the deepest change.

These first comparisons are however made under a specific choice of the number of clusters, given by the DBHT. One could wonder what happens changing this parameter, i.e. moving along the hierarchical structure provided by each clustering method. Let us stress that the DBHT method gives automatically the number of clusters that is instead an adjustable parameter for the other methods. However, DBHT can also be analysed for a varying number of clusters by thresholding over the clustering hierarchical structure. In the following we discuss a set of quantitative analyses that explore the cluster structure at all the hierarchical levels.

4.2.2 Measuring the heterogeneity of clusters size distribution

In the previous Section we have seen that the SL shows a giant cluster that contains more than 90% of stocks, whereas DBHT, CL and k-medoids methods have a more homogenous distribution of cluster sizes. To characterize such differences with a single quantity we can calculate for each clustering the so-called coefficient of variation y [188], defined as:

$$y = \frac{\sigma_S}{\langle S \rangle} \ , \tag{4.1}$$

where σ_S is the standard deviation of clusters size:

$$\sigma_S = \sqrt{\frac{1}{N_{cl}-1}\sum_a (S_a - \langle S \rangle)^2} \ , \tag{4.2}$$

and the normalization factor $\langle S \rangle$ is the average

$$\langle S \rangle = \frac{1}{N_{cl}}\sum_a S_a, \tag{4.3}$$

with S_a being the size of cluster a and N_{cl} the number of clusters. In the limit of homogeneous arrangement of stocks among the clusters (i.e. each cluster has the same number of stocks), we obtain $\sigma_S = 0$ and then $y = 0$. The higher is the degree

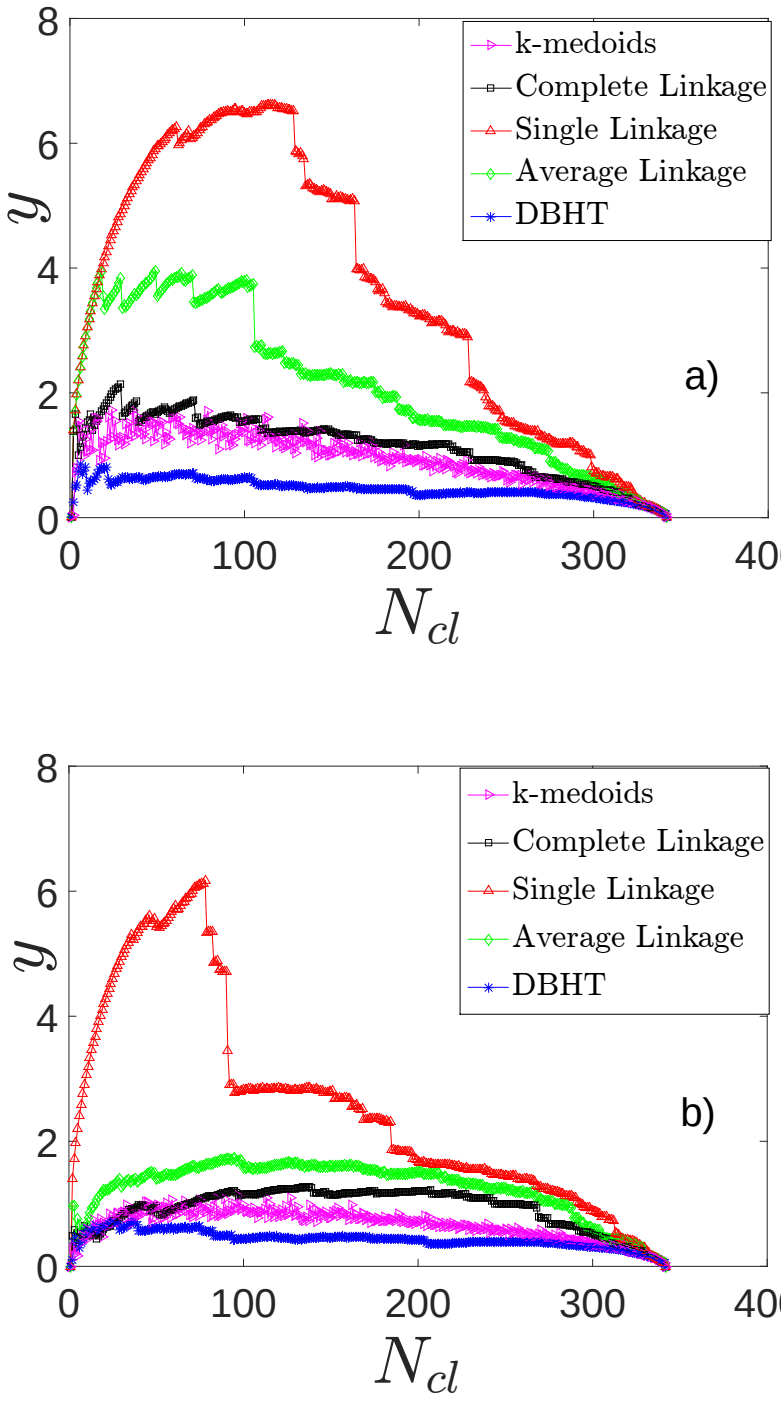

Fig. 4.11 **Demonstration that different clustering methods show different degrees of disparity in the clustering structure**. The disparity measure y is shown for clusterings at different hierarchical levels as a function of N_{cl} in the dendrograms, for a) non-detrended log-returns and b) detrended log-returns.

of heterogeneity in the distribution of sizes, the higher is σ_S and therefore y. In the following we have used the expression "disparity" to refer to y, in order to stress the fact that we use it as a measure of heterogeneity in clusters' sizes.

Through y we can quantitatively compare the five clustering methods in terms of homogeneity in the clusters size distribution. We perform this analysis by varying the number of clusters N_{cl}, to have a more complete picture and explore the complete dependence structure. With the hierarchical clustering methods we obtain this by cutting the dendrograms at different levels of distance; for the k-medoids, for which no dendrogram is present, N_{cl} is simply an input parameter of the algorithm.

In Fig. 4.11 we show, for each clustering method, how the disparity measure varies with N_{cl}. Fig. 4.11 a) shows the non-detrended case, Fig. 4.11 b) the detrended case. As we can see the SL provides the higher disparity in both cases, regardless of N_{cl}; then the AL, CL and k-medoids follow. The DBHT values are below all of them, which means the DBHT provides a more homogeneous community assignment at any level of the correlation hierarchy. Moreover, in the non-detrended case the SL and the AL show the highest values of disparity for N_{cl} in the interval 50-100. The CL and DBHT have instead a flatter pattern, with the highest values occurring for lower values of N_{cl}. Looking at the detrended case in Fig. 4.11 b), the removal of the market mode smooths also the pattern of the AL, whereas the SL is even sharper. Overall, subtracting the market mode makes the clusterings more homogeneous, suggesting that the largest clusters that emerged in SL and AL in the non-detrended case are associated to the market mode dynamics.

The algorithms of SL and AL are indeed expected to be more sensitive to the market mode. In the iterative procedure that generates the SL dendrogram, for instance, the correlation between two new clusters is defined as the maximum correlation between elements of the first cluster and elements of the second one: since the most part of correlation in the market is due to the market mode [68] such an algorithm is likely to force many clusters to join the cluster made of the most influencial stocks in the

market, resulting in a giant cluster and high disparity. The AL is less sensitive to this effect as the inter-clusters correlation is defined as the average of correlations; for the CL the minimum correlation is chosen, resulting -unsurprisingly- in the lowest value of disparity. For what concerns the DBHT it is probably the topology of PMFG, which is more structured and clustered than the MST, to provide a lower sensitivity to the market mode dynamics.

We can conclude that, from the point of view of the disparity measure, the analysed clustering methods provide quite different structures at any level of the dendrograms. The DBHT yields the most homogeneous clustering, whereas the SL displays the highest levels of disparity.

4.2.3 Retrieving economic information: Adjusted Rand Index

In this Section we quantify the amount of economic information retrieved by the clustering methods by measuring the similarity between clustering and ICB. Such similarity have been computed with the Adjusted Rand Index ($\mathscr{R}_{adj}$) [99], which is a tool conceived to compare different clusterings (or community structures) on the same set of items [189]. An industrial sector classification is indeed nothing but a partition in communities of the N stocks. Therefore we can take the similarity between clustering and industrial sector classification as a proxy for the information detected by the clustering method. In particular, given two community structures on the same set of items, $\mathscr{R}_{adj}$ returns a numerical value equal to 1 for identical clusterings and to 0 for completely independent clusterings (namely, whose overlapping is consistent with a random overlapping hypothesis).

The idea behind this index is to calculate the number of pairs of objects that are in the same cluster in both clusterings, and then to compare this number with the one expected under the hypothesis of independent clusterings. Specifically, and following the notation of [189], let us call X the set of the N objects (stocks, in our case). Let us call $Y = \{Y_1, ..., Y_k\}$ a clustering which is a partition of X into communities which

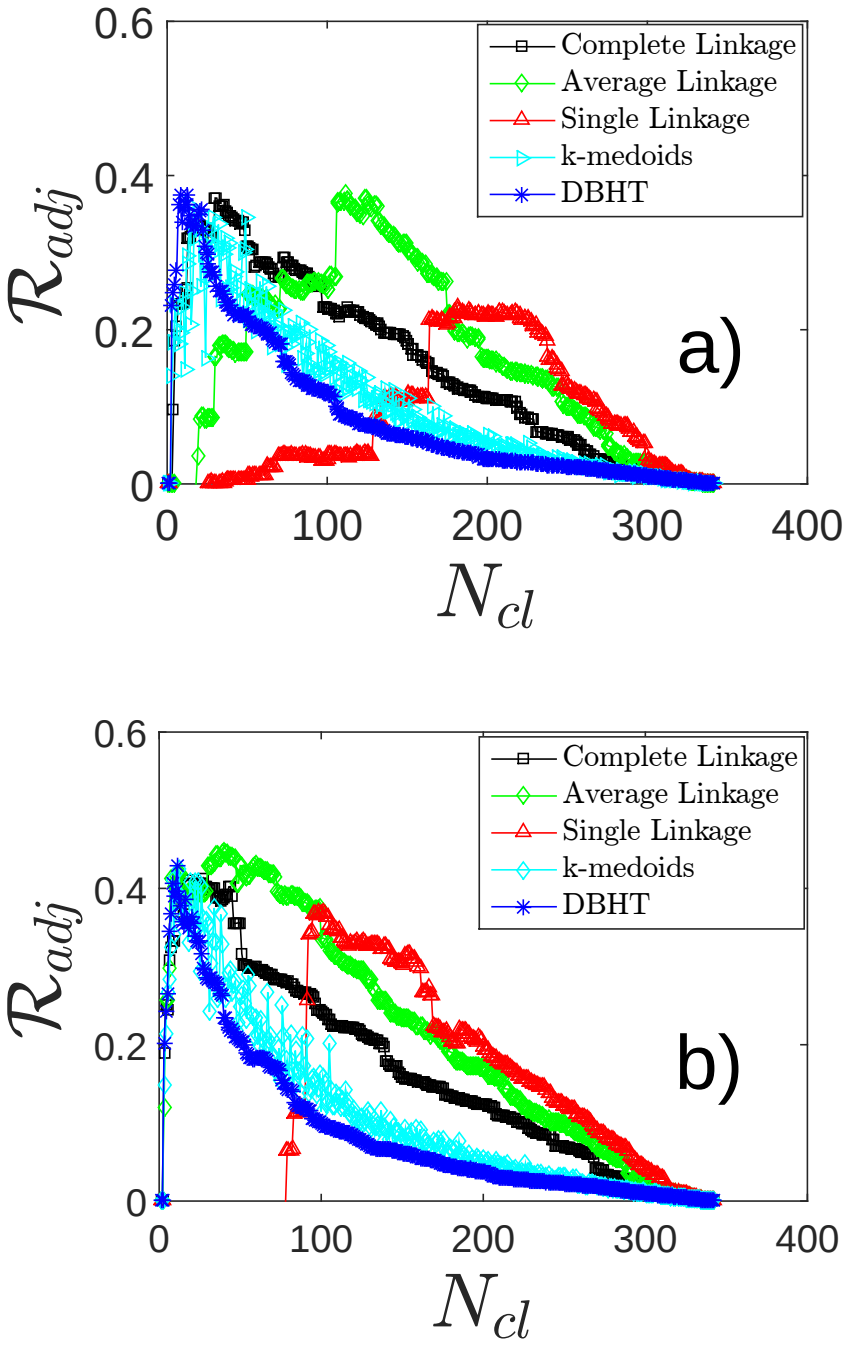

Fig. 4.12 **Demonstration that different clustering methods retrieve to different degrees the ICB industries**. The Adjusted Rand Index $\mathscr{R}_{adj}$ between clustering and ICB industries is shown for different number of clusters N_{cl}. In a) correlations are calculated on non-detrended log-returns, in b) are calculated on detrended log-returns. The vertical dashed line shows the value ($N_{cl} = 10$) correspondent to the actual number of ICB industries.

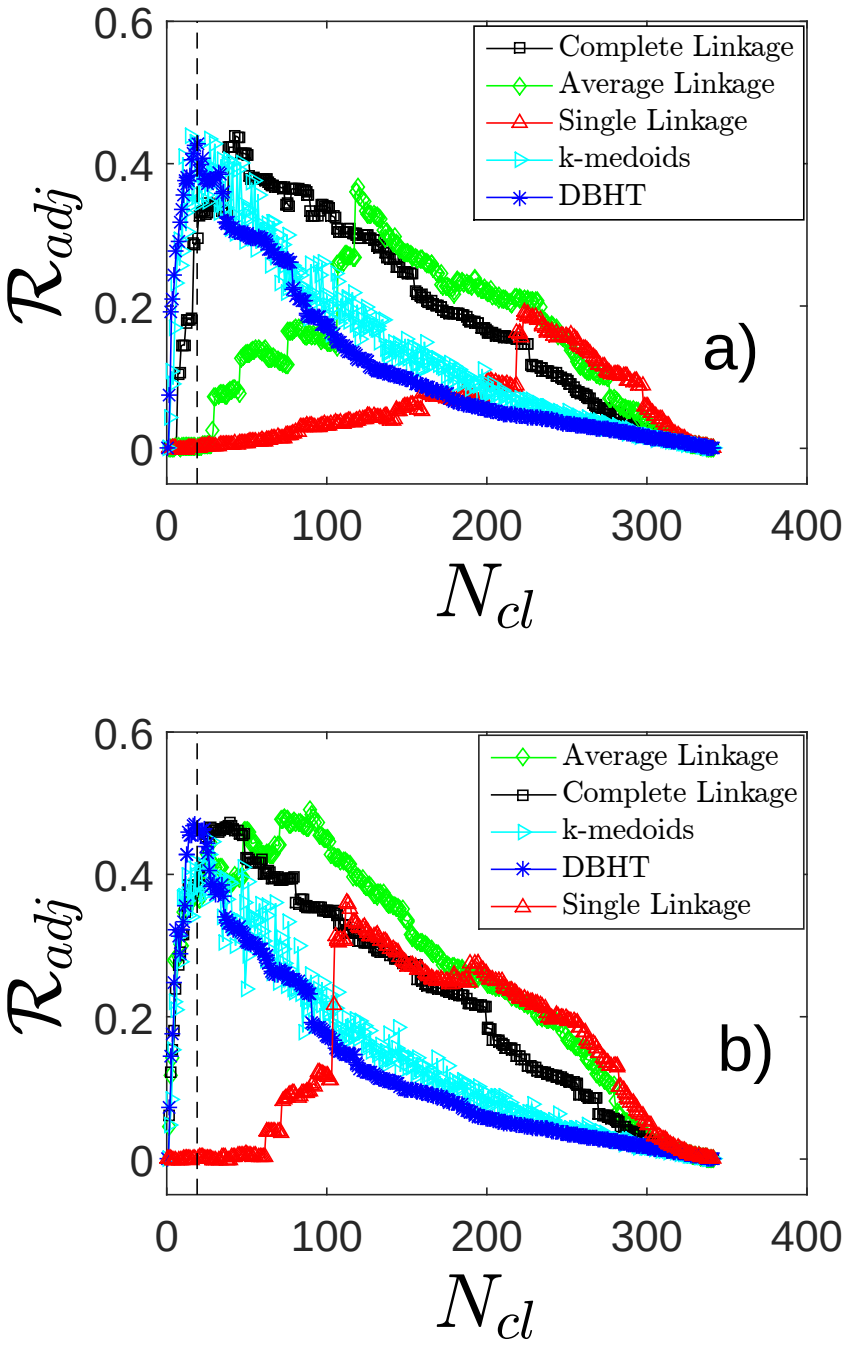

Fig. 4.13 **Demonstration that different clustering methods retrieve to different degree the ICB supersectors**. The Adjusted Rand Index $\mathscr{R}_{adj}$ between clustering and ICB industries is shown for different number of clusters N_{cl}. In a) correlations are calculated on non-detrended log-returns, in b) are calculated on detrended log-returns. The vertical dashed line shows the value ($N_{cl} = 19$) correspondent to the actual number of ICB industries.

are non-empty disjoint subsets of X such that their union equals X: $X = Y_1 \cup ... \cup Y_k$ [189]. Let us also consider another different clustering Y', containing l clusters. We call "contingency table" the matrix $M = \{m_{ij}\}$ with coefficients

$$m_{ij} \equiv |Y_i \cap Y'_j|, \tag{4.4}$$

i.e. the number of objects in the intersection of clusters Y_i and Y'_j . Let us call a the number of pairs of objects that are in the same cluster both in Y and in Y', and b the number of pairs that are in two different clusters in both Y and Y'. Then the Rand Index is defined as the sum of a and b, normalized by the total number of pairs in X:

$$\mathscr{R}(Y,Y') \equiv \frac{2(a+b)}{N(N-1)} = \sum_{i=1}^{k}\sum_{j=1}^{l}\binom{m_{ij}}{2}. \tag{4.5}$$

We then use, as null hypothesis associated to two independent clusterings, a generalized Hypergeometric distribution [9]. The Adjusted Rand Index is then defined as the difference between the Rand Index and its mean value under the null hypothesis, normalized by the maximum that this difference can reach:

$$\mathscr{R}_{adj}(Y,Y') \equiv \frac{\sum_{i=1}^{k}\sum_{j=1}^{l}\binom{m_{ij}}{2} - t_3}{\frac{1}{2}(t_1+t_2) - t_3} \ , \tag{4.6}$$

where:

$$t_1 = \sum_{i}^{k}\binom{|Y_i|}{2} \ , \ t_2 = \sum_{j}^{l}\binom{|Y'_j|}{2} \ , \ t_3 = \frac{2t_1t_2}{N(N-1)} \ . \tag{4.7}$$

We have $\mathscr{R}_{adj} \in [-1,1]$, with 1 correspondent to the case of identical clusterings and 0 to two completely uncorrelated clusterings. Negative values instead show anti-correlation between Y and Y' (that is, the number of pairs classified in the same way by Y and Y' is less than what expected assuming a random overlapping between the two clusterings).

As in the disparity analysis, we have measured $\mathscr{R}_{adj}$ by varying the number of clusters N_{cl}. In Figs. 4.12 and 4.13 we show $\mathscr{R}_{adj}(N_{cl})$ as a function of N_{cl}, by using ICB industries and supersectors respectively. Figs. 4.12 a) and 4.13 a) refer to non-detrended log-returns, whereas Figs. 4.12 b) and 4.13 b) correspond to detrended log-returns. Vertical dashed line in the graphs identifies the values $N_{cl} = 10$ (Fig. 4.12) and $N_{cl} = 19$ (Fig. 4.13), that is the number of ICB industries and supersectors.

Let us focus on the non-detrended case in Figs. 4.12 a) and 4.13 a) first. Maximum values of $\mathscr{R}_{adj}$ - $\mathscr{R}^*_{adj}$ from now on - are reached at different values of N_{cl}, depending on the clustering method. In particular DBHT, CL and k-medoids maxima are shifted towards low values of N_{cl} with respect to AL and SL. This means that economic information is gathered at different levels of the hierarchical structure depending on the clustering method. Moreover, DBHT, CL and k-medoids maxima $\mathscr{R}^*_{adj}$ are slightly higher in the comparison with ICB supersectors than with industries, whereas for SL and AL the reverse is true. Overall, the SL structure displays the lowest similarity, whereas the other methods have comparable $\mathscr{R}^*_{adj}$ around 0.4: this value could therefore indicate the actual amount of economic information present in the dependence structure.

The detrended case in Figs. 4.13 b) and 4.13 b) shows several differences. We notice first of all that $\mathscr{R}^*_{adj}$ increase for all the methods. The natural interpretation for this is that the market mode, driving all the stocks regardless of their industry and supersector, hides to some extent the economic structure [68]. Secondly, also maxima locations on the N_{cl} axis change. This effect is in particular strong for AL and SL, whose maxima are shifted towards left, closer to the other methods. In the detrended case all methods but SL provide clusterings more similar to supersectors than industry.

4.2.4 Retrieving economic information: ICB overexpression

The Adjusted Rand Index provides an overall measure of similarity between the clustering partition and the industrial classification [99]. In order to analyse to what extent each industrial sector is retrieved by the clusters we must look at the stocks in common

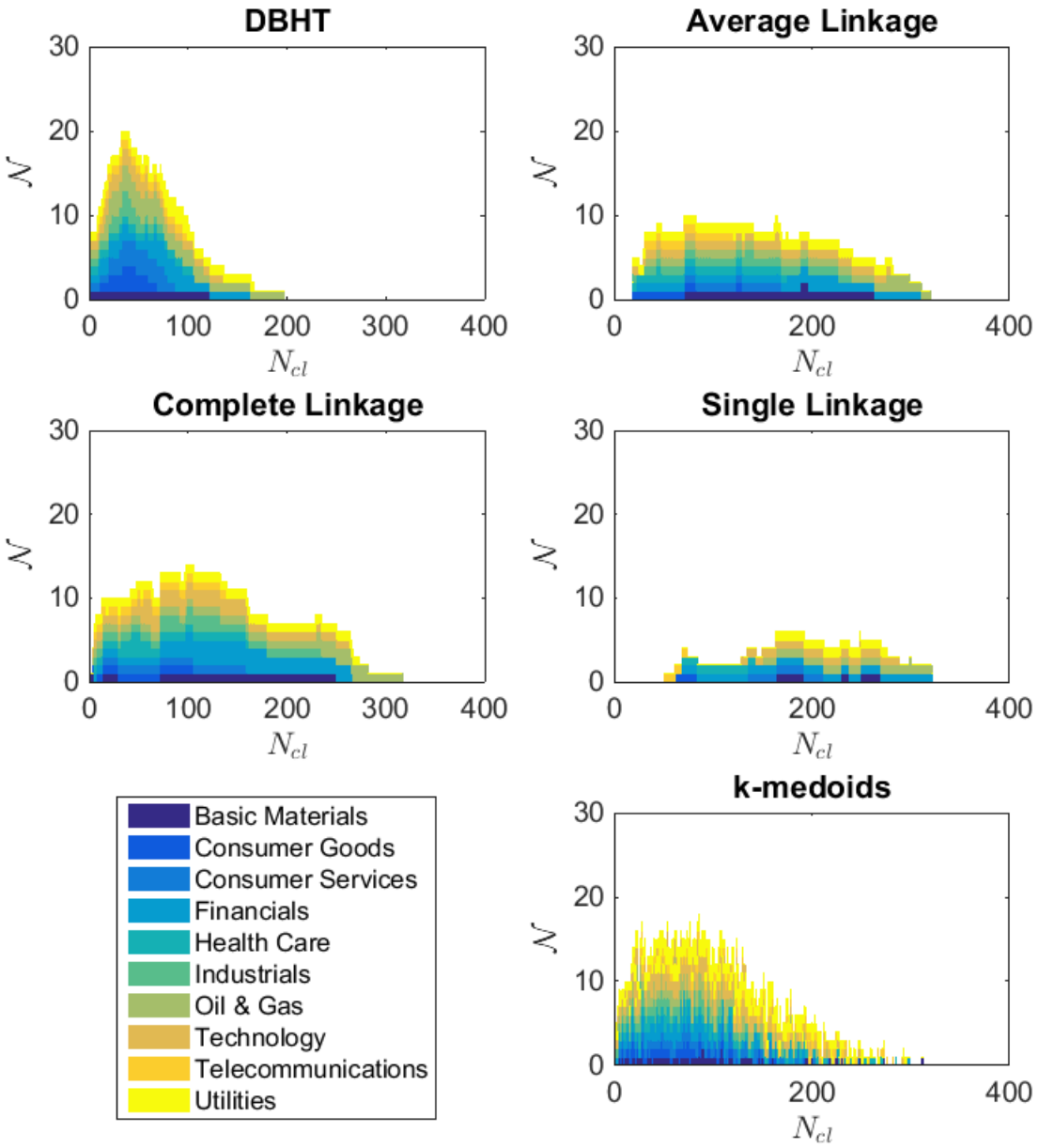

Fig. 4.14 **Amount of ICB information retrieved by the clustering methods, in terms of ICB industries overexpressions**. Each bar graph shows, varying the number of clusters N_{cl}, how many times ($\mathcal{N}$) an ICB industry is overexpressed by a cluster according to the Hypergeometric hypothesis test (i.e., number of null-hypothesis tests being rejected). Each colour shows the number of overexpressions for each ICB industry. In graphs a)-e) the results for DBHT, AL, CL, SL and k-medoids clustering are shown respectively. The correlations are calculated on log-returns.

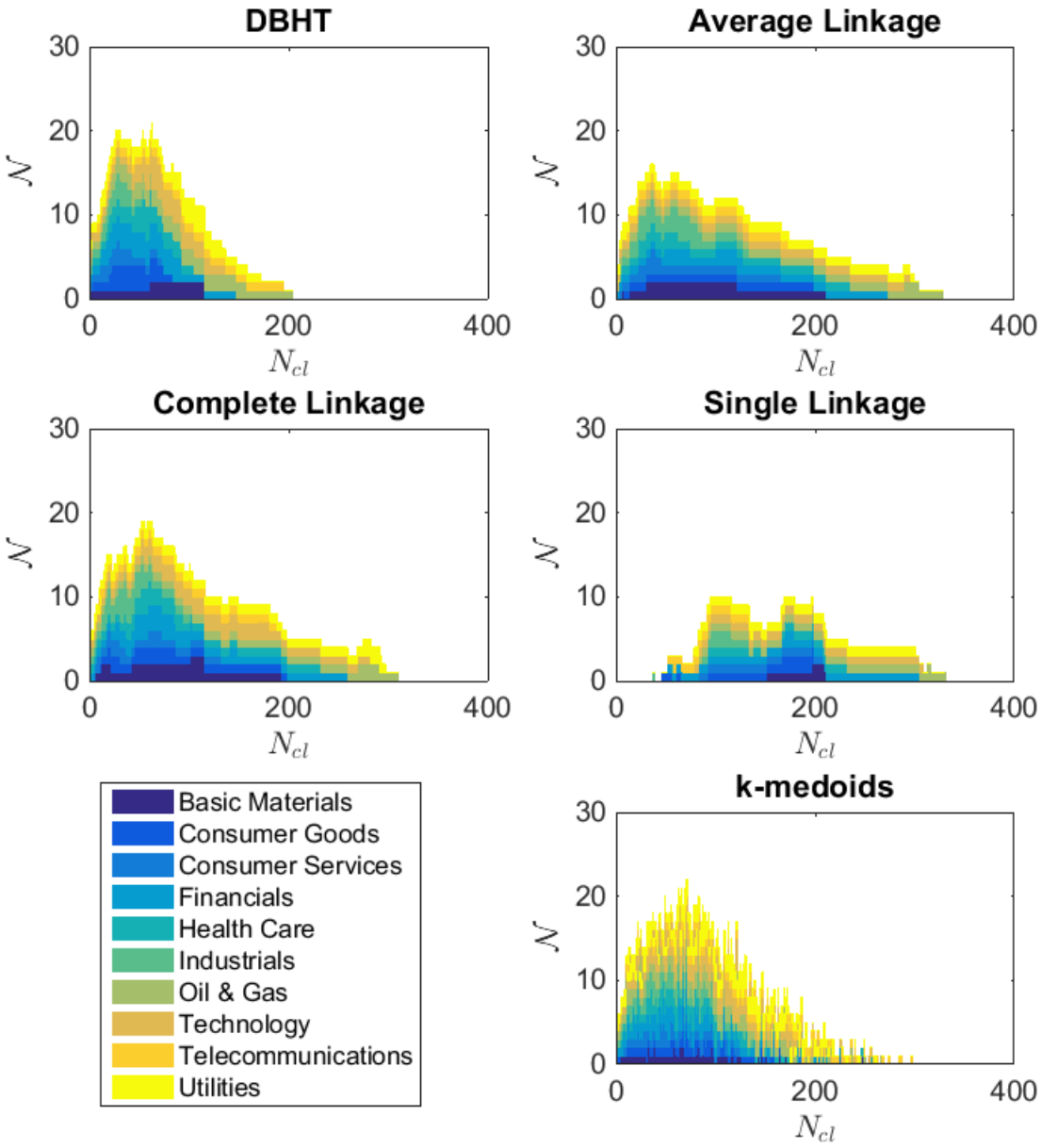

Fig. 4.15 **Amount of ICB information retrieved by the clustering methods, in terms of ICB industries overexpressions. Detrended log-returns case**. Each bar graph shows, varying the number of clusters N_{cl}, how many times ($\mathcal{N}$) an ICB industry is overexpressed by a cluster according to the Hypergeometric hypothesis test (i.e., number of null-hypothesis tests being rejected). Each colour shows the number of overexpressions for each ICB industry. In graphs a)-e) the results for DBHT, AL, CL, SL and k-medoids clustering are shown respectively. The correlations are calculated on log-returns detrended of the market mode.

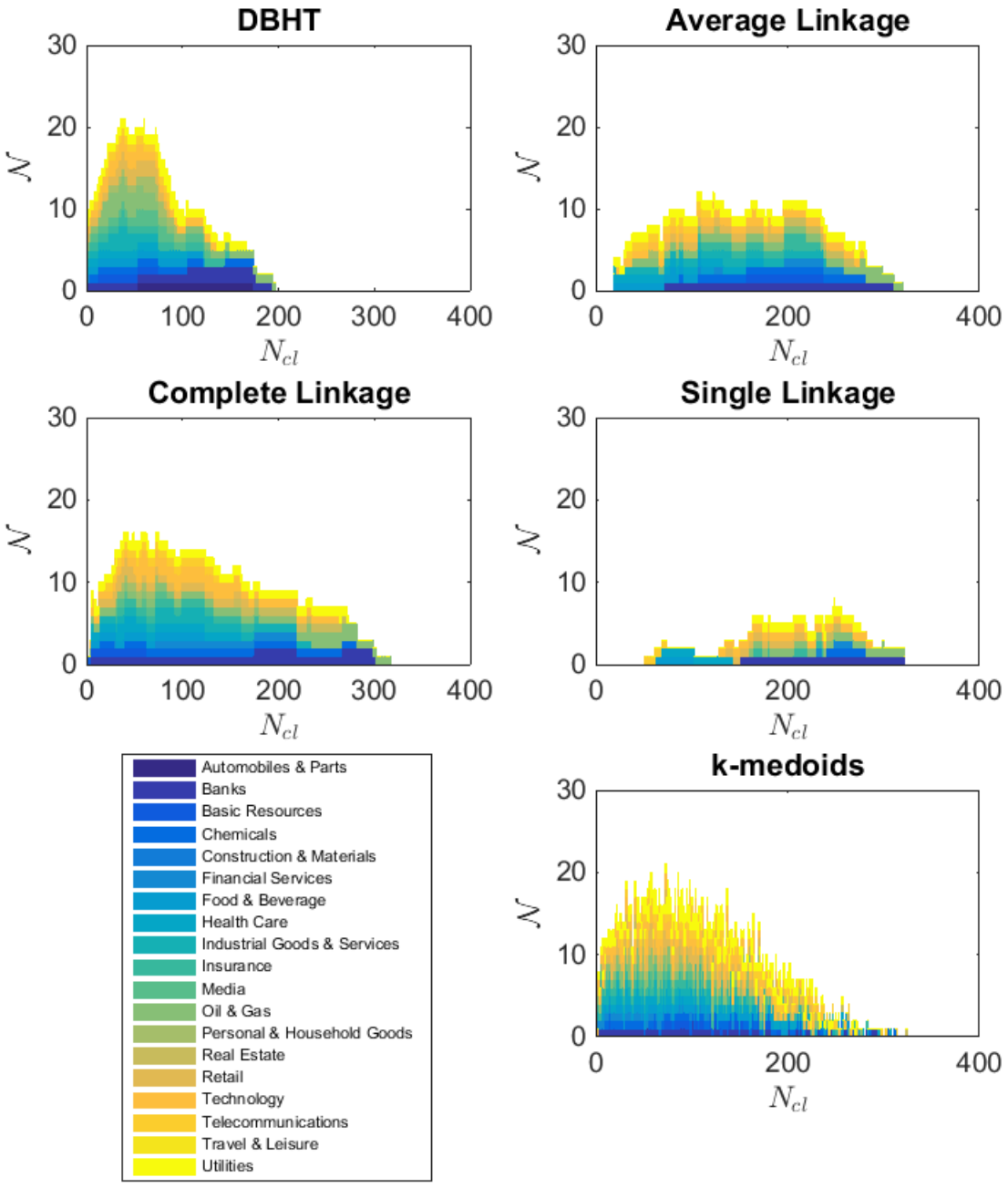

Fig. 4.16 **Amount of ICB information retrieved by the clustering methods, in terms of ICB supersectors overexpressions**. Each bar graph shows, varying the number of clusters N_{cl}, how many times ($\mathcal{N}$) an ICB supersector is overexpressed by a cluster according to the Hypergeometric hypothesis test (i.e., number of null-hypothesis tests being rejected). Each colour shows the number of overexpressions for each ICB supersector. In graphs a)-e) the results for DBHT, AL, CL, SL and k-medoids clustering are shown respectively. The correlations are calculated on log-returns.

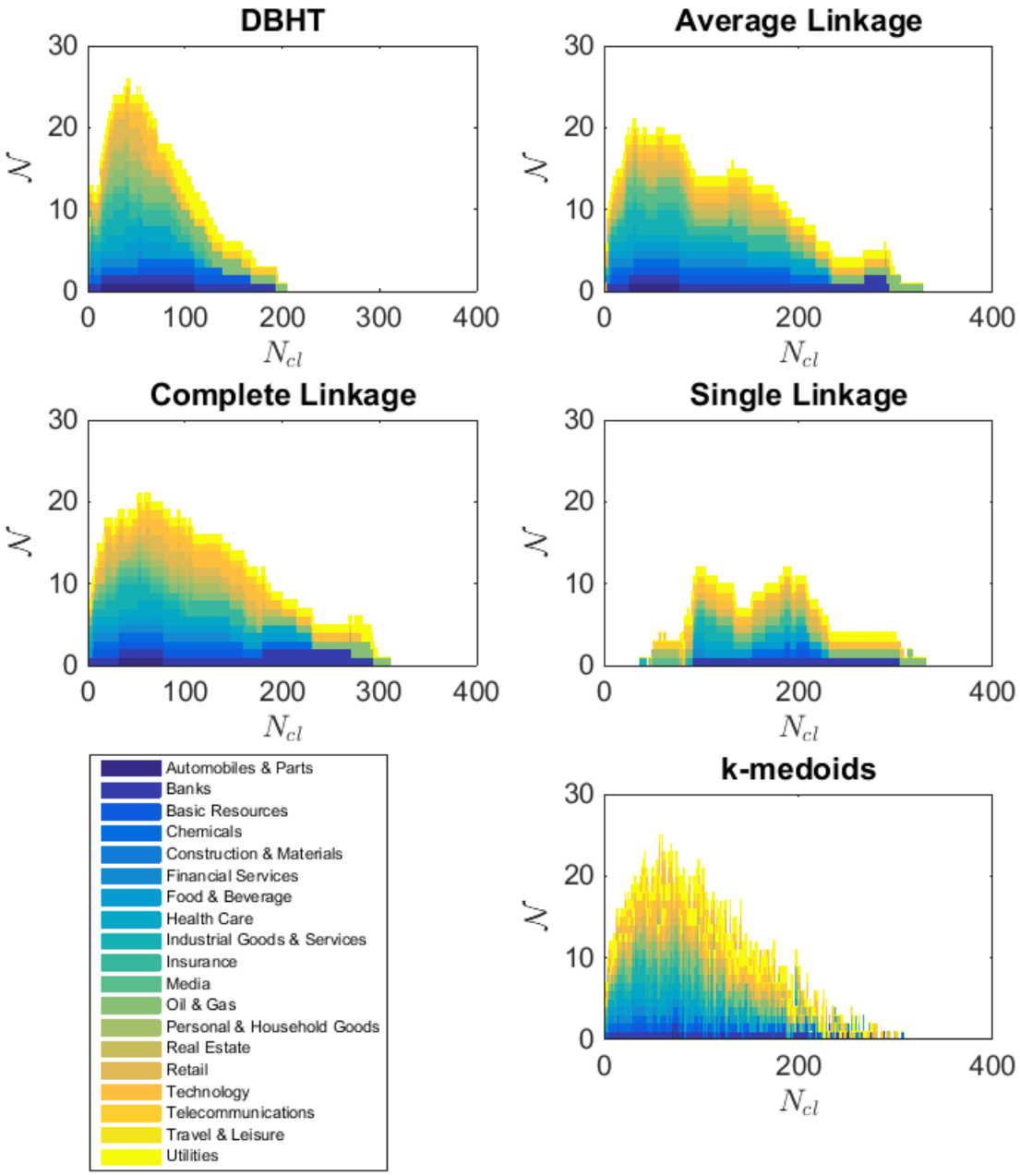

Fig. 4.17 **Amount of ICB information retrieved by the clustering methods, in terms of ICB supersectors overexpressions. Detrended log-returns case**. Each bar graph shows, varying the number of clusters N_{cl}, how many times ($\mathcal{N}$) an ICB supersector is overexpressed by a cluster according to the Hypergeometric hypothesis test (i.e., number of null-hypothesis tests being rejected). Each colour shows the number of overexpressions for each ICB supersector. In graphs a)-e) the results for DBHT, AL, CL, SL and k-medoids clustering are shown respectively. The correlations are calculated on log-returns detrended of the market mode.

between each pair industry/cluster (and supersector/cluster). If the percentage of stocks in common is sensitively higher than what expected from a random overlapping of communities we say that the cluster overexpresses the industry (or supersector).

To quantify such overexpression we use a statistical one-tail hypothesis test [9], where the null hypothesis is the Hypergeometric distribution [96] which describes the probability that by random chance two communities of given size have in common k objects over a total of N [9, 98]. In particular, let us call Y_i a cluster in our clustering and Y'_j a sector. We want to verify whether Y_i overexpresses Y'_j. If k is the number of stocks in common between Y'_j and Y_i, and $|Y_i|$, $|Y'_j|$ are the cardinalities of the cluster and the sector respectively, then the Hypergeometric distribution is [98]:

$$P(X=k) \;=\; \frac{\binom{|Y'_j|}{k}\binom{N-|Y'_j|}{|Y_i|-k}}{\binom{N}{|Y_i|}} \quad . \tag{4.8}$$

This is the null hypothesis for the test: to be distinguishable by a random overlap the number k of stocks in common must be significantly different from a random overlap and therefore $P(X=k)$ must be small. If $P(X=k)$ is less than the significance level, then it is said that the test is rejected. If the test is not rejected, then it means that we cannot reject the hypothesis that the k stocks in Y_i coming from a sector Y'_j are picked up just by chance, without any preference for that sector. If instead the test is rejected, we conclude that the cluster Y_i overexpresses the sector Y'_j.

We have performed this hypothesis test for each pair of cluster and ICB industry/supersector. This amounts to $\frac{1}{2}N_{cl}N_{ICB}$ tests in total, where N_{ICB} is the number of ICB industries or supersectors. We have then counted the number $\mathscr{N}$ of hypothesis tests that are rejected, i.e. that shows significant overlapping between a cluster and a ICB industry or supersector. The higher $\mathscr{N}$ is, the higher is the economic information contained in the correlation clustering. The advantage with respect to $\mathscr{R}_{adj}$ is that we are able to distinguish contributions to $\mathscr{N}$ from different ICB industries and supersectors. Indeed in Figs. 4.14 - 4.17 we show $\mathscr{N}$ as a function of N_{cl}, with different colors

showing each industry/supersector contribution to $\mathscr{N}$ (that is, how many times that industry/supersector has been found to be overexpressed by a cluster). We have chosen a significance level for the test equal to 0.01, together with the conservative Bonferroni correction [141, 98] that reduces the significance level to $0.01/(\frac{1}{2}N_{cl}N_{ICB})$.

As we can see, composition and shape of $\mathscr{N}$ depends strongly on the clustering method. The DBHT and k-medoids reach the highest $\mathscr{N}$ values, that means they retrieve the highest amount of economic information. Moreover their corresponding $\mathscr{N}$ shape is quite peaked, quickly dropping to low values for high N_{cl} values. The economic information is therefore gathered in a narrow region of N_{cl} for these two methods. On the contrary, the three Linkage methods display lower overexpression and are flatter, indicating that their economic information is spread along the N_{cl} axis. When the market mode is detrended though these differences reduce and in particular CL and AL take a more peaked shape. Overall, and consistently with the previous results, removing the market mode increases $\mathscr{N}$ for all methods.

For what concerns industrial and supersectorial composition, we can see that the DBHT, k-medoids, CL and AL show a quite homogeneous composition, with almost each ICB supersector overexpressed. The SL instead shows a much less rich composition, with no more than 6 overexpressed supersectors and industries simultaneously even at the maximum level of total overexpressions. In terms of composition the k-medoids exhibits a high level of noise against N_{cl}, whereas the hierarchical methods are much more stable.

Finally, it is worth noticing that there is a change in the composition at different values of N_{cl}, and that similar patterns can be found across the four hierarchical clustering methods (for the k-medoids no clear patterns can be found, because of the higher level of instability of the method). There are industries and supersectors that tend to become overexpressed for low values of N_{cl} and then disappear at intermediate values: this is the case of Automobiles & Parts, Telecommunications, Insurance and Financial Services. Others are instead more persistent, appearing along all the x-axis: Utilities,

Technology, Health Care and Oil & Gas. The most persistent is the latter, that is still overexpressed when all the others are not expressed anymore. We can then conclude that not only the ICB partition is hidden at different levels in the dendrograms depending on the clustering method, but also different ICB supersectors are retrieved at different levels. This is probably due to the different degrees of correlation within different ICB industries and supersectors.

4.3 The dynamical evolution of the clustering structure

Here we present a dynamical analysis of the DBHT clustering in the 15 years ranging from 1 January 1997 to 31 December 2012. We have selected the set of overlapping time windows described in subsection 2.4.4 of Chapter 2 ($n = 100$ time windows of length $L = 1000$ trading days) and used the weighted version of the Pearson estimator (Eq. 2.19 in subsection 2.4.4) in order to mitigate excessive sensitiveness to outliers in remote observations.

In Fig. 4.18 a) the number of DBHT clusters obtained for each time window is shown, both for non-detrended log-returns (red circles) and detrended log-returns (blue squares). For the first case the number of clusters ranges between 6 and 19, for the second case the range is 14-26. The dashed lines are the values correspondent to the clustering obtained using the entire period 1997-2012 as time window.

As observed previously, the number of clusters in the non-detrended case is sistematically lower than the detrended case. Moreover, an overall decreasing trend characterizes the non-detrended values and makes them go below the corresponding dashed line; this decreasing pattern is not present in the detrended case, that however stays below the correspondent dashed line the most of the times either. It is interesting also to analyse the evolution of the disparity y, introduced in Eq. 4.1, over the period. In Fig. 4.18 b) we show y for each time window, both for the non-detrended and detrended case. Again the dashed lines are the values for the all period clusterings. In the non-detrended case

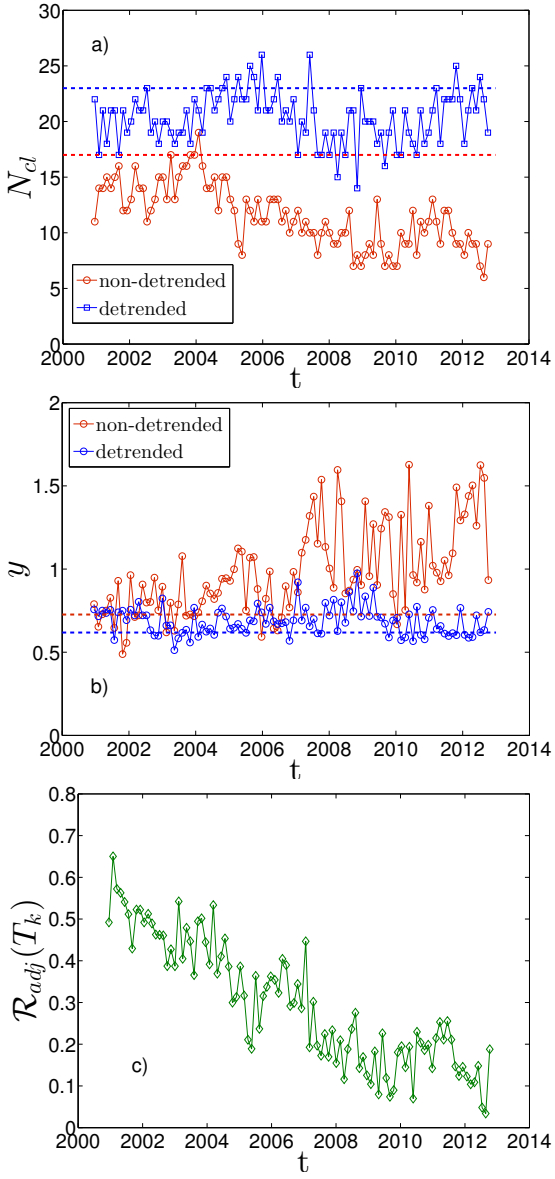

Fig. 4.18 **Dynamical evolution of the DBHT clustering**. Each plot refers to 100 moving time windows of length 1000 trading days. Specifically, in graph a) we plot the number of DBHT clusters, N_{cl}, for both log-returns non-detrended (red circles) and detrended by the market mode (blue squares), whereas the two dashed horizontal lines are the N_{cl} values obtained by taking the largest time window of 4026 trading days. Overall the non-detrended case shows a decreasing trend. In graph b) it is shown the disparity measures, y, again for the two sets of DBHT clustering (red dots non-detrended, blue dots detrended), the dashed horizontal lines being the y values from the 4026 length time window. In the non-detrended case the 2007 marks a transition to higher and more volatile values of y. Finally in graph c) it is shown the Adjusted Rand Index, $\mathscr{R}_{adj}$, measured at each time window between the detrended and non-detrended clusterings. A steady decreasing trend is evident.

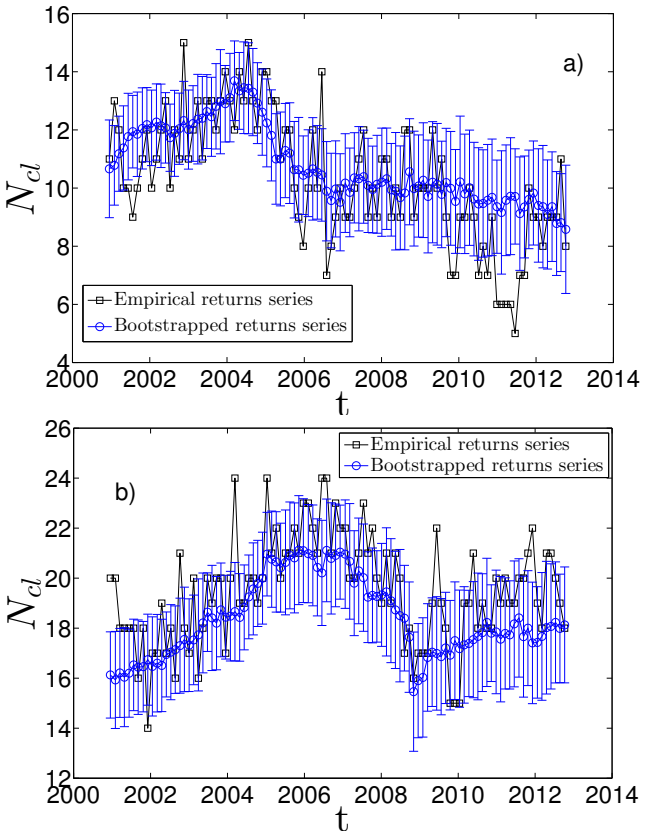

Fig. 4.19 **Test of robustness for the dynamical DBHT clustering**. a) Number of clusters N_{cl} as a function of the time t: the black squares correspond to the DBHT clusterings obtained by using the empirical (non-detrended) log-returns, the blue dots are the average over the 100 N_{cl} given by the 100 replica correlation matrices (see text for further details). The bar errors in the blue dot plot is the standard deviation calculated among the same set of 100 N_{cl}. As one can see the empirical N_{cl} is quite robust against the bootstrapping test. b) Same plot as in a), but by using detrended log-returns.

we see an overall increasing trend, especially after the 2006; an analysis of the sizes distribution show that in this period the largest cluster contains up to 240 stocks (70% of total number of stocks); moreover, from 2006 on we observe also a much higher fluctuation in the values. This behaviour is of interest since it concerns the overall influence of the market mode on the correlation structure, with higher y indicating a stronger influence of the market mode that tends to gather all stocks in one cluster. Indeed, in the detrended case we find that subtracting the market mode makes the

increasing trend disappear. Overall the disparity values decrease and stay closer to the dashed line, without significant pattern apart from some fluctuations.

In order to better understand the relation between the DBHT clusterings obtained with detrended and non-detrended log-returns, we have also performed a dynamical Adjusted Rand Index analysis. Now we compare no longer the clustering and the ICB partition, but the two clusterings (non-detrended and detrended) at each time window. In Fig. 4.18 c) the Adjusted Rand Index between the two sets of DBHT clusters is shown. Interestingly, it appears a steady decreasing trend that drives the similarity from relatively high values (about 0.7) to values close to zero, indicating complete uncorrelation between the two clusterings. We can therefore conclude that the influence of the market mode has increased remarkably over the last 15 years, making the detrended clustering structure more and more different from the non-detrended one. This observation would not have been possible without the clustering analysis, since from the preliminary dataset measures (see e.g. Fig.2.16) it is not evident any constant pattern either in the average return or in the average correlation.

In order to test the sensitiveness of the DBHT clustering to the statistical noise, inevitably present in every correlation estimate, we have performed a bootstrapping test [190, 191] to the equity dataset. This is a non-parametric method able to estimate the error of a given estimator, without relying on any assumption about the true distribution [14]. A detailed description of the bootstrapping method is provided in Appendix B. In Fig. 4.19 we show the result of a dynamical bootstrapping, performed over all the 100 time windows. At each time window we have run $B = 100$ permutations. The blue points are the average number of clusters over the B replicas, whereas the error bars are the standard deviations calculated over the same sample. The black squares are the empirical numbers of clusters yielded by the DBHT. The plot a) is by using non-detrended log-returns, the plot b) by using detrended log-returns. The plot of empirical number of clusters is slightly different from what we have shown in Fig. 4.18 a) because for this bootstrapping analysis we did not use exponential smoothing for the

correlations, but only bare correlations. The exponential smoothing, indeed, creates an asymmetry among the points in each time series that makes the bootstrapping test inapplicable. From the plot we can observe that the method is statistically robust, with the most of empirical points within one standard deviation from the mean of replicas. More importantly, the mean of replicas follows the general trend of the empirical points; namely, the decreasing trend in the market mode case, and the drop after the 2007-2008 credit crunch in the detrended case.

4.3.1 Dynamically retrieving the industrial sectors

Let us here investigate the relation between industrial classification and clustering under a dynamic perspective. To this end we here perform the previous dynamical analysis by considering the set of 100 overlapping time windows T_k and calculating for each of them the Adjusted Rand Index $\mathscr{R}_{adj}(T_k)$ between clustering and ICB partition. Since $\mathscr{R}_{adj}(T_k)$ varies with the chosen threshold and N_{cl}, we select at every time the N_{cl} that maximizes $\mathscr{R}_{adj}(T_k)$; the numbers that we report are these maximum values and account therefore for the maximum ability of the clustering methods to retrieve the ICB.

In Figs. 4.20 a)-e) we show the results for each of the five clustering methods, using returns with market mode. Interestingly, all of them show a decreasing trend in time. On average, the DBHT and CL display the highest similarity with industrial classfication, whereas the Single Linkage the lowest. This is consistent with what found in the static analyses. We have also highlighted in the graphs the major events that affected the stock market in the last 15 years. It can be observed that different clustering methods are affected in different ways by these events. Indeed, if the 2007-2008 credit crunch crisis and the following recession is evident in all methods as a significant drop in the similarity, other events such as 11/09/2001 or the 2002 stock market downturn appear only in the Single and Average Linkage plots. In particular the 2002 downturn drives a steep decrease in the similarity of SL and AL, that stay at low values until the end of 2005. For DBHT, Complete Linkage and k-medoids instead these events do not seem

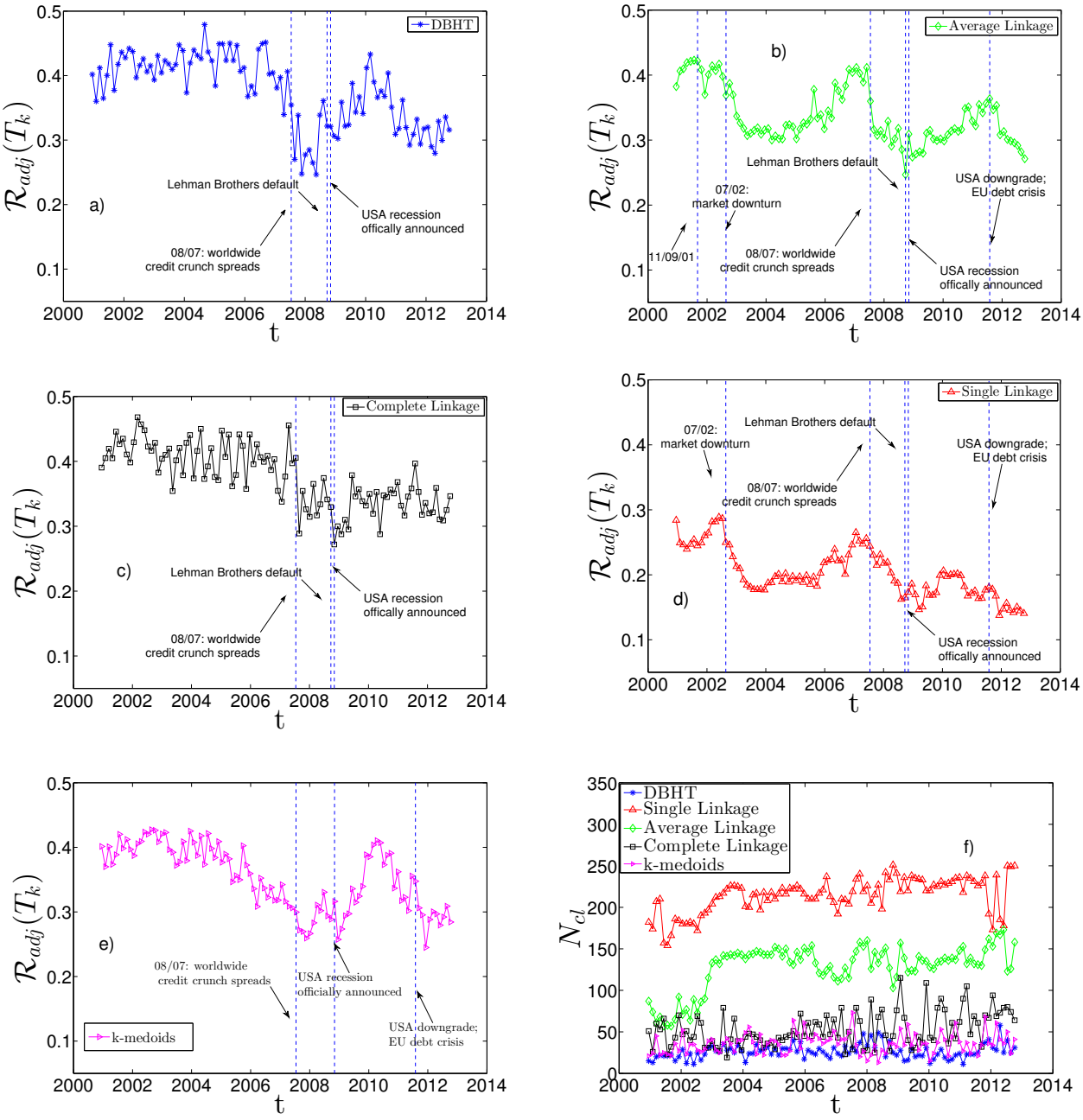

Fig. 4.20 **Dynamical evolution of the similarity between clustering and ICB**. It is shown the Adjusted Rand Index, $\mathscr{R}_{adj}$, calculated at each time window T_k ($k = 1,...,n$) between clustering and ICB partition, for a) DBHT, b) AL, c) CL, d) SL and e) k-medoids method. At each time window the number of clusters, N_{cl}, has been chosen in order to maximize the $\mathscr{R}_{adj}$ itself: in f) we plot these N_{cl} values for each clustering method. It is evident as the maximum similarity clustering-ICB is reached at different hierachical levels depending on the clustering method. The correlations are calculated on log-returns with market mode.

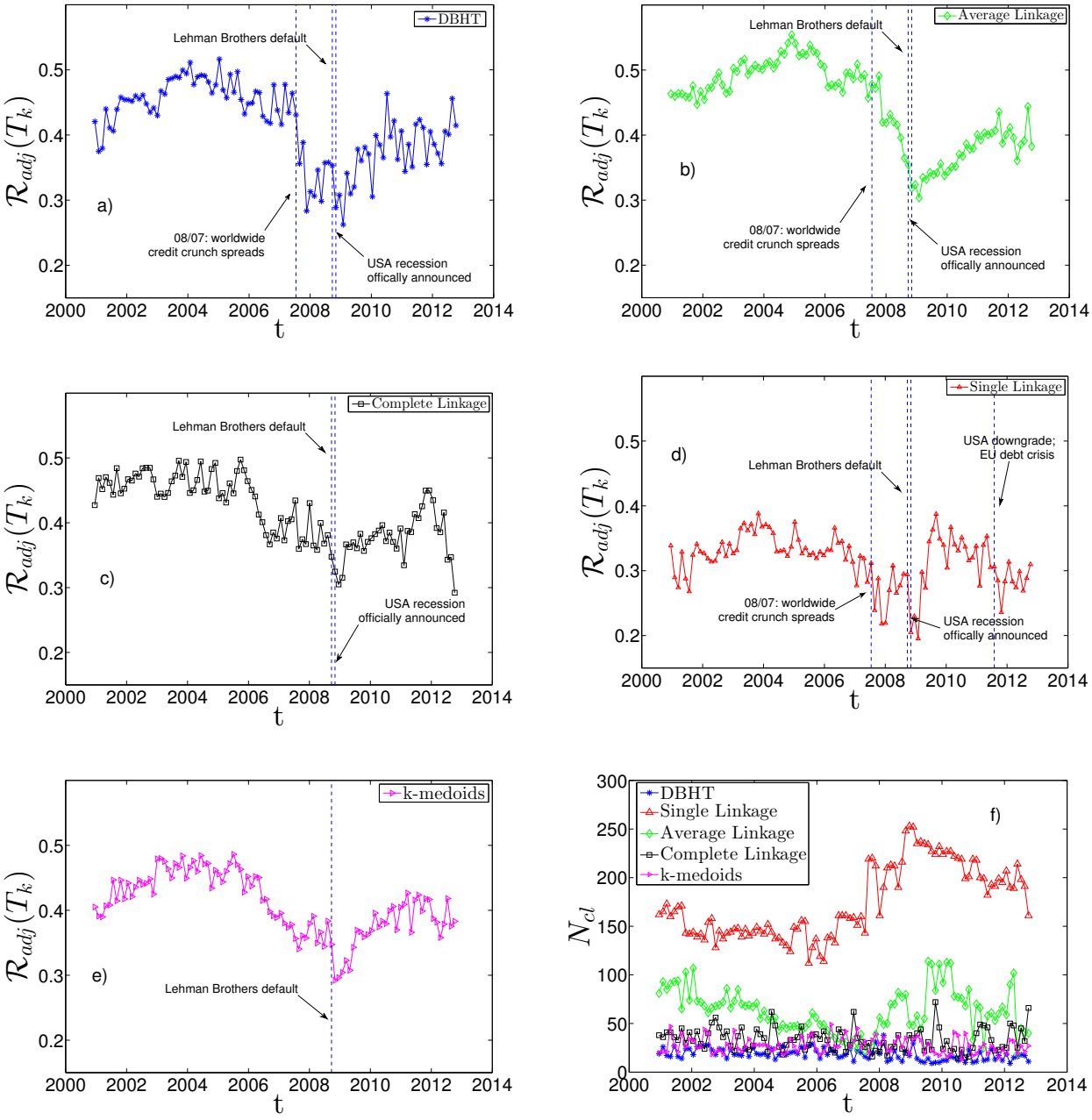

Fig. 4.21 **Dynamical evolution of the similarity between clustering and ICB, with detrended log-returns**. a)-f): same graphs as in Fig. 4.20, but by using correlations on detrended log-returns.

to affect the similarity in a noticeable way compared to the statistical fluctuations. This observation points out that the DBHT, CL and k-medoids are more robust than SL and AL against exogenous events in their ability to retrieve an economic information as the industrial classification. Nonetheless, there are differences also among DBHT, CL and k-medoids: in particular in the period following the 2008 crisis, DBHT and k-medoids

show a peak that does not appear with CL. Moreover, for the k-medoids the drop in similarity seems to begin more than one year before the 2007. All these features have non-trivial implications for both portfolio optimization and systemic risk evaluation.

Fig. 4.20 f) shows the number of clusters N_{cl} that maximizes, in each time window T_k, the Adjusted Rand Index shown in the previous plots. As we can see, N_{cl} for SL is always the highest, followed by AL, CL, k-medoids and DBHT. This is consistent with what we found in the static analysis in Section 4.2: different clustering methods "retain" the industrial information at different levels of the hierarchy. SL and AL, that yield higher N_{cl} (i.e., lower levels in the hierarchy), are also the methods that show the lowest level of similarity with industrial classification and the highest degree of disparity.

In Figs. 4.21 a)-f) we show the same set of plots for the detrended case. The main differences with the non-detrended case are the following:

- the average similarity with the industrial classification rises for all methods; this confirms in the dynamical case what we found for the static case;

- the average N_{cl} is lower for all methods: the absence of market mode "moves" the industrial classification to higher levels of the hierarchy;

- the strong influence of the 11/09/2001 and 2002 downturn on the SL and AL pattern seems to disappear, whereas the 2007-2008 crisis is still evident in all the five methods. This could be explained claiming that the former are global events in the market, whereas the latter exhibits also a "local" dynamics;

- the AL shows the most evident change in the dynamical behaviour, displaying a trend much more similar to the DBHT and CL one. Also in terms of N_{cl}, it shows values closer to DBHT, CL and k-medoids than SL.

4.4 Summary

In this chapter we have presented a set of static and dynamical analyses to empirically quantify the information filtered from correlation matrices by different hierarchical clustering methods. The use of unsupervised learning techniques allowed us to make no assumptions on the returns distribution.

We have analysed the correlations among log-returns of $N = 342$ US stock prices, across a period of 15 years (1997-2012). We have compared five clustering methods: Single Linkage, Average Linkage, Complete Linkage, k-medoids and the DBHT, which has been applied to financial data for the first time here. We have taken the Industrial Classification Benchmark (ICB) as industrial sector partition for the stocks [192]. The degree of similarity with correlation-based communities has been measured by using tools as the Adjusted Rand Index [99] and the hypergeometric hypothesis test [97]. We have focused not only on the communities of asset, but on the entire hierarchies associated to them, covering all the different levels of the hierarchical structures. The dynamical perspective of our study is crucial for applications, in particular for what concerns hedging risk and portfolio optimization: for this reason we have given a particular attention to the effects of financial crises on the hierarchical structures, highlighting differences among the clustering methods.

The clustering methods show remarkably different performances in retrieving the economic information encoded in the ICB, with big dissimilarities even among the Linkage methods. We have suggested that these differences should be connected to different degrees of sensitivity to the market mode dynamics, that in turns are to be ascribed to differences in the methods underlying working principles. Moreover, the economic information appears to be retained at different levels of the hierarchical structures depending on the clustering method. The DBHT and k-medoids methods show the best performances, but the latter seems to be affected by the noise much more than the DBHT and the Linkage methods. The DBHT turns out then to be a good mix between the advantages of the k-medoids and those of the Linkages. The dynamical

analysis has also proved that the methods show different degrees of sensitivity to financial crises. This is again a new result that could give insights into the dynamics of such events, as well as an indication on which clustering method is more robust for financial applications.

We have also performed each analysis on log-returns detrended by the market mode, by following a standard procedure in literature [68, 120]. Interestingly the effect of this detrending is very dissimilar for different methods, with the weakest methods (Average and Single Linkage) improving remarkably their ability to retrieve industrial sectors. In general the detrending increases the degree of economic information that the clustering methods retrieve. It also makes the distribution of cluster sizes more homogeneous (suggesting that the high heterogeneity in SL and AL must be due to the market mode dynamics), as well as more stable against time. Finally, the dynamical analyses have shown that the clustering structure reveals peculiar patterns over the financial crisis showing an increasing dominant role of the market mode over the period 1997-2012, implying an increase of the non-diversifiable risk in the market.

In Chapter 5 we investigate further the dynamical evolution of the dependence structure through clustering analysis. In particular we will extend the approach introduced here to tackle the problem of non-stationarity in financial correlation.

Chapter 5

Evolution of correlation-based networks and clusters tracking

In this chapter we investigate the persistence in time of the dependence structure analysed in Chapter 4. Our approach is again model-free, unlike traditional tests for stationarity, and is based on correlation-based networks. Chiefly, we track the evolution of individual clusters using statistical hypothesis tests, and we analyse their changing composition in terms of industrial sectors. We find evidence of strong non-stationarity and we discuss the implication for risk diversification strategies.

Part of the results and analyses presented in this chapter has been published in the paper "Risk diversification: a study of persistence with a filtered correlation-network approach" in 2015 [193].

5.1 Introduction

A way to reduce financial risk is diversifying investments taking positions in assets that are historically anti-correlated or uncorrelated, reducing in this way the probability that all assets loose value at the same time. However, the applicability of these approaches relies on the implicit assumption that the relevant features of the dependence structure observed in the past have persistent significancy into the future. This is not always the

case, as discussed in Chapter 2: it is generally accepted in the literature that financial correlations are non-stationary [38].

In order to address this issue, in this chapter we have taken the dynamic analyses of Chapter 4 a step further. Here we aim at estimating the degree of non-stationarity in the market correlation by using PMFG networks and the associated DBHT clustering. In this context persistence translates into a measure of similarity among communities in a network, for which network-theoretic tools should be used. The advantage over traditional tests of stationarity is again the model-free nature of our approach, that does not require any assumption on the log-returns distribution.

The original contributions of this chapter are the following:

- We introduce a new measure of similarity between dependence structures at different periods, based on the Adjusted Rand Index and the DBHT clustering. We use this measure to quantify and study the rate of change of the dependence structure, especially during the 2007-2008 crisis. We find that this structure displays a phase transition in correspondence with the crisis.

- We track the evolution of the DBHT clusters through a set of hypergeometric hypothesis tests. This allows us to investigate the changing composition of each cluster, revealing peculiar patterns in correspondence with the financial crisis. In particular we find that industrial sectors have become less useful for risk diversification.

- We investigate the PMFG evolution by computing network metrics such as degree and clustering coefficient at different time windows. The time series we obtain in this way are then analysed through correlograms and power-law fits, revealing the existence of long-term memory patterns in the evolution of the PMFG topology.

The rest of the chapter is organized as follows: in Section 5.2 we investigate the persistence of the dependence structure: in particular in Subsection 5.2.1 we focus on the global structure, using DBHT and the Adjusted Rand Index as a measure of persistence,

whereas in Subsection 5.2.2 we study the evolution of each single DBHT cluster; in Section 5.3 we investigate the dynamics in terms of correlation-based networks topology, by studying metrics such as nodes' degree through time series analysis tools.

5.2 Persistence and transitions: dynamical analysis of DBHT

In this section we present a set of dynamical analyses aiming to characterize the persistence and evolution of the dependence structure through the associated DBHT clustering. We have studied the set of equities that we have introduced in Chapter 2; namely, daily prices of $N = 342$ US stocks, covering the period from January 1997 to December 2012. We have then selected the set of overlapping time windows described in Section 2.4.4 of Chapter 2: $n = 100$ overlapping time windows of length $\theta = 1000$ trading days, with 30 trading days shift between adjacent time windows.

Then in each time window we have computed the weighted Pearson coefficient defined in Eq. 2.19. Given the richer and more robust clustering associated with detrended log-returns, we have calculated Pearson correlation coefficients on residuals $c_i(t)$ defined in Eq. 2.17 in Chapter 4.

Firstly we characterize the persistence in terms of economic information expressed by the Adjusted Rand Index $\mathscr{R}_{adj}(T_k)$ (see subsection 4.2.3) between ICB partition and DBHT clustering at time window T_k. Unlike the dynamical analysis in Chapter 4.3 we here explore all of ICB hierarchical levels (namely the subsectors, sectors, supersectors and industries), by computing a different $\mathscr{R}_{adj}(T_k)$ for each level. In Fig. 5.1 a) we therefore show the evolution in time of $\mathscr{R}_{adj}(T_k)$ between DBHT clusters and ICB industries, supersectors and subsectors (for sake of simplicity we do not plot the sectors data that are very close to supersectors values). The ICB information at all levels shows a remarkable drop during the 2007-08 financial crisis, to be partially recovered from 2010 onwards. Interestingly before the crisis the industry, supersector

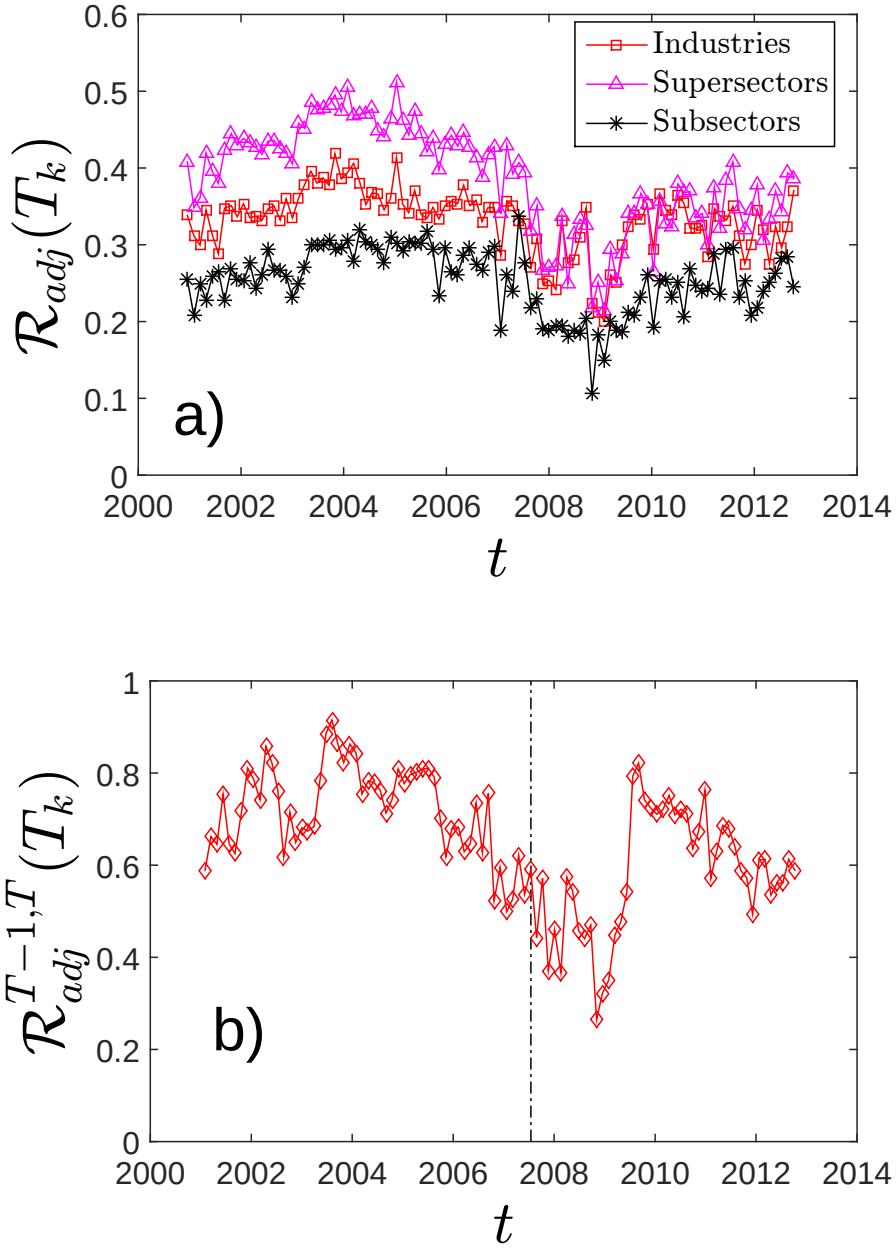

Fig. 5.1 **Dynamical evolution of the DBHT clustering**. Each plot refers to 100 moving time windows (T_k) of length 1000 trading days and shifted of 30 days. a) Amount of economic information retrieved by DBHT clustering in terms of similarity between clustering and ICB partitioning calculated by using the Adjusted Rand Index, $\mathscr{R}_{adj}$. A drop at the outbreak of crisis appears. Over the post-crisis years the economic information is less than in the pre-crisis period and differences among different ICB levels are less evident. b) Persistence of DBHT clustering in time, measured as the Adjusted Rand Index between two adjacent clusterings. The financial crisis is characterized by very low levels of persistence.

and subsector lines were clearly distinct (with ICB supersectors showing the highest similarity with DBHT, followed by industries and subsectors) whereas in the crisis and post-crisis periods they display much closer values. Therefore from the crisis onwards the correlation clustering is no longer able to distinguish between different levels of ICB: this might indicate that this industrial classification is becoming a less reliable benchmark to diversify risk.

The Adjusted Rand Index can also be used as a tool for analysing the persistence of DBHT clustering by measuring the index between two clusterings at two adjacent time windows (we denote $\mathscr{R}_{adj}^{T-1,T}(T_k)$ such a quantity). This gives a measure of local persistence: a drop in the index value indicates decreasing similarity between adjacent clusterings, and therefore lower persistence. In Fig. 5.1 b) we plot $\mathscr{R}_{adj}^{T-1,T}(T_k)$ against time. We can observe that the clustering persistence changes remarkably over time, dropping in particular with the outbreak of financial crisis and recovering in 2010. It is worth pointing out that the drop during the crisis starts earlier than the actual outbreak of it (August 2007, dashed vertical line): this could highlight a possible use of clustering persistence as tool to forecast systemic risk. Notably, in the time period 2010-2012 we observe again a steady decreasing trend. Interestingly the pattern of persistence appears to be related to the similarity between clustering and ICB, with periods of higher persistence characterized by higher amount of economic information.

However the drawback of $\mathscr{R}_{adj}^{T-1,T}(T_k)$ as a measure of persistence is that at any time it only provides information on the persistence with respect the previous, adjacent time window. It tells nothing about long-term robustness of each clustering. To investigate this aspect we discuss in the next section a set of analyses that evaluate the persistence of each clustering at each time providing therefore a more complete picture.

5.2.1 A map of structural changes

To investigate the long-term persistence of each clustering we have calculated for each time window the Adjusted Rand Index between the correspondent clustering and the

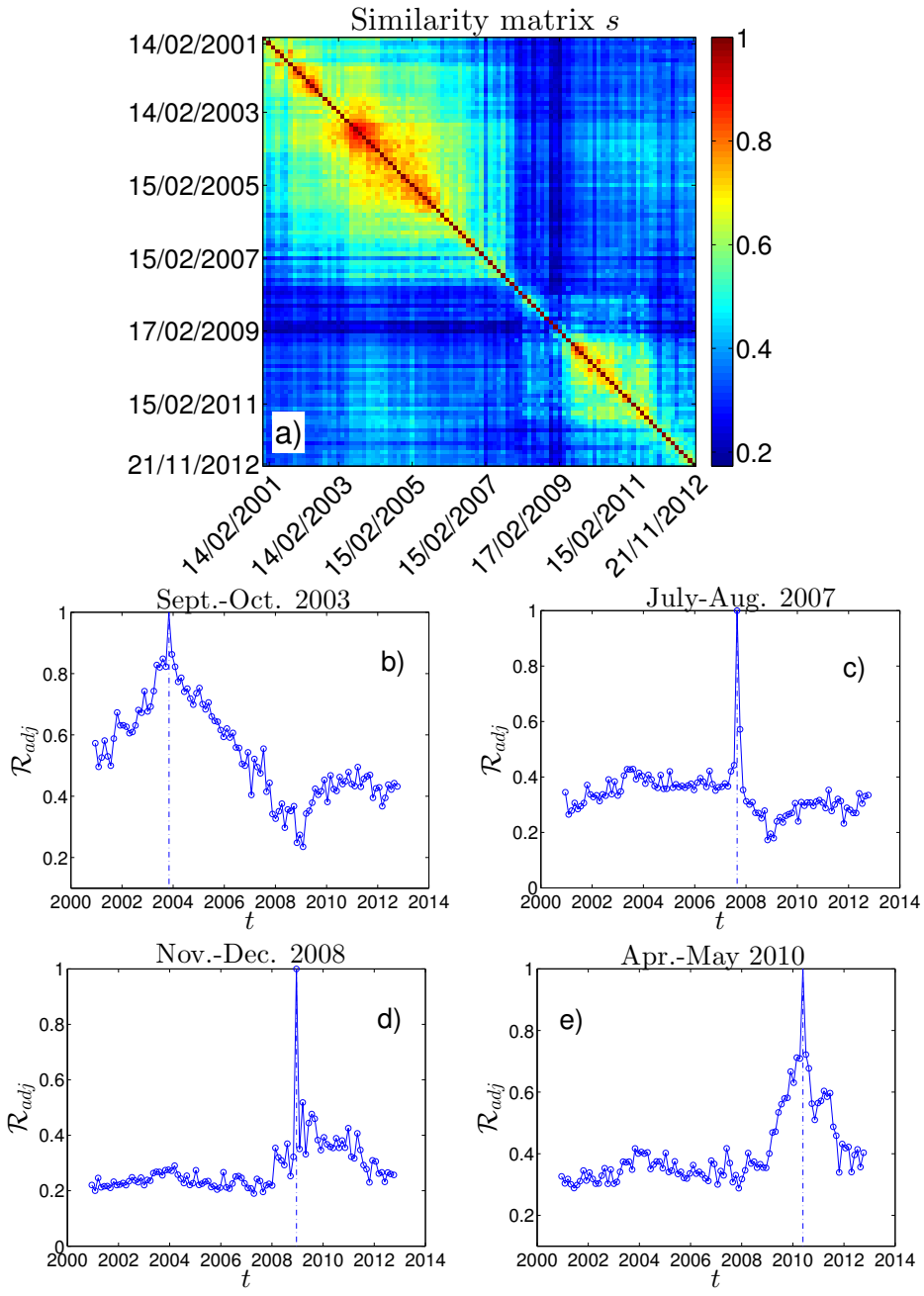

Fig. 5.2 **Persistence analysis based on clustering**. a) Similarity matrix s showing the temporal evolution of the correlation-based DBHT clustering. Each entry $s(T_a, T_b)$ is the Adjusted Rand Index between clustering X_a and X_b at time window T_a and T_b respectively (Eq. 5.1): higher values indicate higher similarity. The matrix displays two main blocks of high intra-similarity, one pre-crisis and the other one post-crisis. The years 2007-2008 fall between these two blocks and display very low similarity with any other time window, revealing an extremely changeable structure. Figures b)-e) show the patterns of similarity for four sample time windows (i.e. four sample rows of the similarity matrix): during the crisis the decay of similarity becomes much faster than in the pre and post-crisis periods.

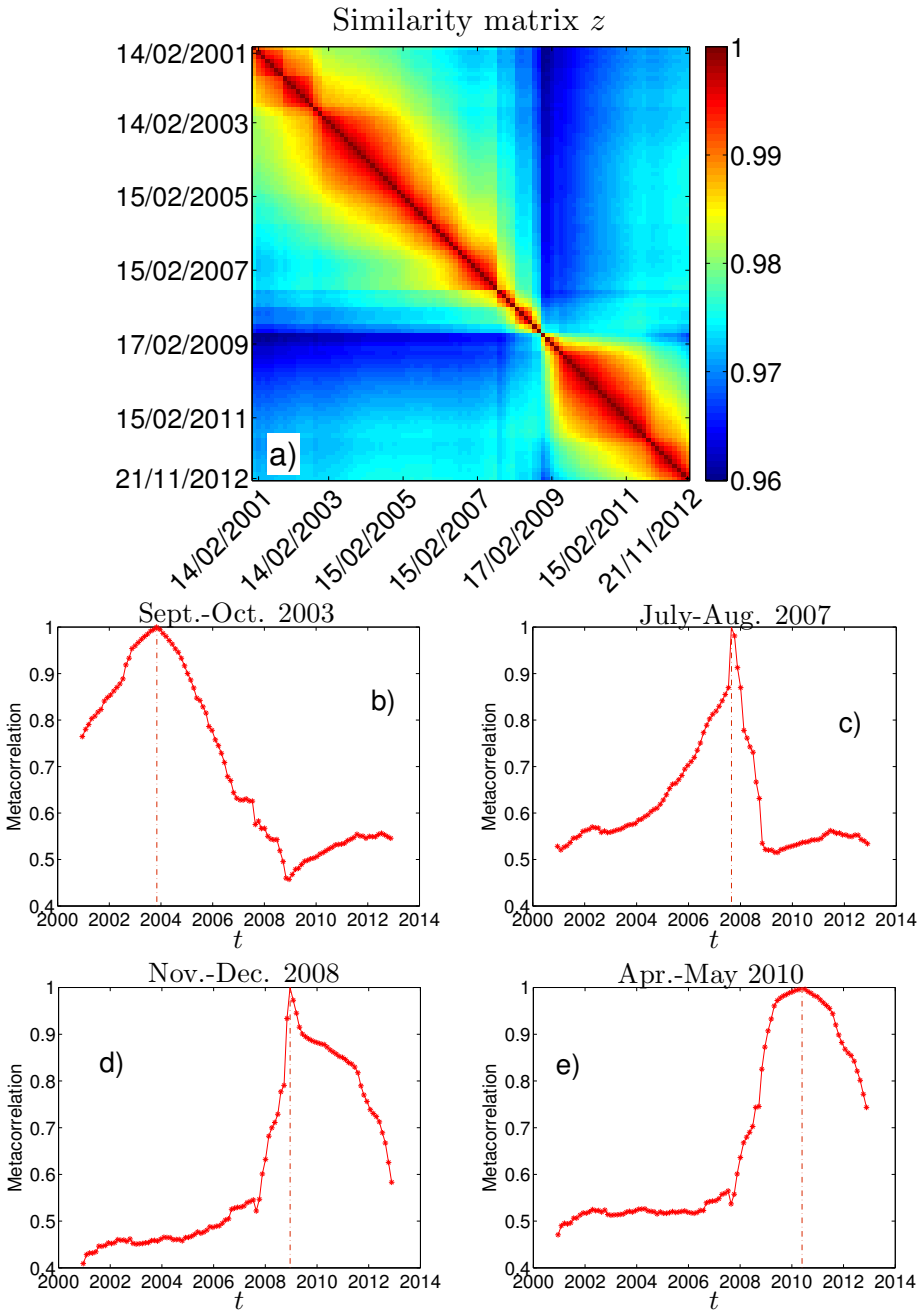

Fig. 5.3 **Persistence analysis based on metacorrelation**. a) Similarity matrix z showing the temporal evolution of correlation matrices. Each entry $z(T_a, T_b)$ is calculated as correlation among correlation matrices at time windows T_a and T_b (Eq. 5.2): higher values indicate higher similarity. Figures b)-e) show the patterns of similarity for four sample time windows: the decay during the crisis years is much less steep than for the corresponding plot in Fig. 5.2.

clustering at any other time: the result is summarized in the (symmetric) similarity matrix s:

$$s(T_a, T_b) = \mathscr{R}_{adj}(X_a, X_b) \ , \tag{5.1}$$

where X_a and X_b are the DBHT clusterings at time windows T_a and T_b respectively. The matrix s for the dataset is shown in Fig. 5.2 a). We observe two main blocks, the first pre-crisis and the other post-crisis, within which we find high similarity among clusterings. The two blocks show very low mutual similarity (upper right corner/lower left corner of the matrix). The first block begins losing its compactness in 2007, and the second one quite quickly at the beginning of 2009: between these two periods the outbreak of financial crisis displays a series of extremely changeable clusterings, that do not show similarity with any other time window.

To better highlight these changes of regime we plot in Figs. 5.2 b) - e) four time rows from matrix s, taken as examples of persistence behaviour during the pre-crisis (September-October 2003, b)), crisis (July-August 2007, the outbreak of crisis, and November-December 2008, the aftermath of Lehman Brothers default, c) and d)) and post-crisis period (April-May 2010, e)). Each point in the plot is the Adjusted Rand Index between the clustering identified by the dashed vertical line and all the other clusterings at other time windows, both in the past and the future. In the pre-crisis period b) the similarity displays quite a slow decay both forward and backward in time: the original clustering has still 60% of similarity with the 17th time window forward/backward in time. During the crisis, in c) and d), the pattern changes drastically: the similarity drops by 70-80% in few months both backward and forward in time. The two stages of crisis reveal also some differences: while in the early crisis period c) the similarity with pre-crisis clusterings is higher than with the post-crisis ones, in the post Lehman Brothers period d) the situation is reversed. Finally, the post-crisis period e) shows a partially recovered persistence, although not at the same levels of the 2003 pattern.

One could wonder whether these structural changes highlighted by the clustering analyses can be detected directly by studying the original, unfiltered correlation matrices.

To check this we introduce an alternative measure of similarity among different time windows that does not make any use of clustering, namely the correlation calculated between the coefficients of two correlation matrices (metacorrelation). This measure is:

$$z(T_a,T_b)=\frac{\langle\rho_{ij}(T_a)\rho_{ij}(T_b)\rangle_{ij}}{\sqrt{[\langle\rho_{ij}^2(T_a)\rangle_{ij}-\langle\rho_{ij}(T_a)\rangle_{ij}^2][\langle\rho_{ij}^2(T_b)\rangle_{ij}-\langle\rho_{ij}(T_b)\rangle_{ij}^2]}}\ , \tag{5.2}$$

where $\rho_{ij}(T_a)$ is the correlation between stocks i and j at time window T_a and $\langle...\rangle_{ij}$ is the average over all couples of stocks i,j. In [39] an alternative measure has been introduced to identify the possible states of a financial market. In Fig. 5.3 we report the matrix $z(T_a,T_b)$ and four representative time rows, corresponding to the same four time windows chosen in Fig. 5.2. We can observe that metacorrelation is indeed able to identify the two pre-crisis and post-crisis time blocks, but shows also a smaller, intermediate block during the 2007-2008 crisis with a relative high intra-similarity. This is different from what we have observed in the clustering based matrix s, where the time windows during the crisis are quite dissimilar even from each other. Moreover the pre-crisis and post-crisis blocks in z display higher intra-similarity than s, especially over the post-crisis years. All these differences can be appreciated looking at the four z time rows in Figs. 5.3 b)-e): even if in the crisis time windows c) and d) a faster decay of similarity can be observed, the decay is much less steep than the corresponding clustering plot (Figs. 5.2 c) and d)). Moreover the post-crisis window e) recovers completely the high pre-crisis level of persistence, unlike the clustering case in Fig. 5.2 e).

Therefore it seems that metacorrelation and clustering analysis depict slightly different dynamics for the market dependence structure. In particular the clustering based matrix s reveals higher non-stationarity during the crisis and the post-crisis period. The instability of correlation during crises has been recently observed in [194]: however in that work the result relies on a specific choice for the multivariate distribution of returns, whereas our analyses are model independent.

5.2.2 Clusters composition evolution

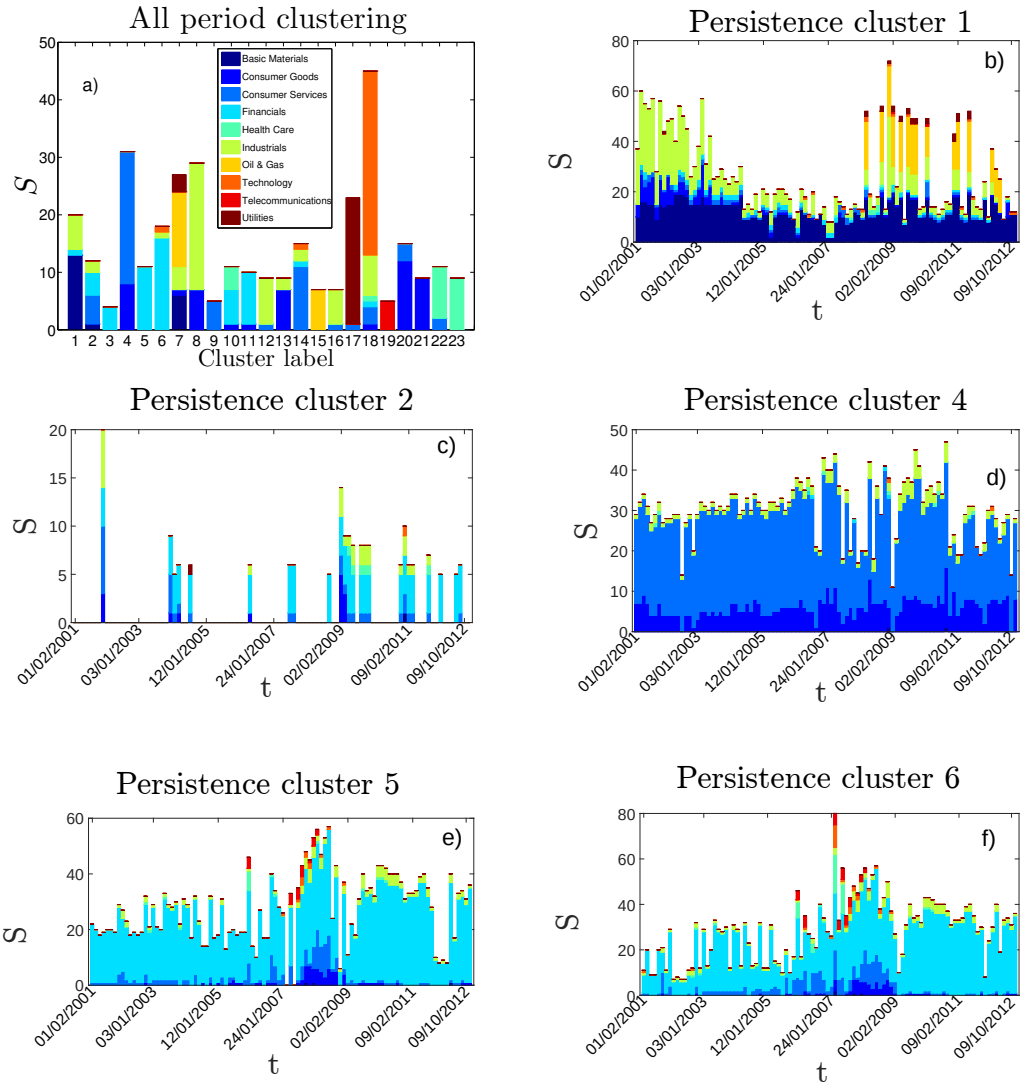

Fig. 5.4 **Clusters dynamical composition (part 1)**. a) Clusters composition of DBHT clusters obtained by calculating detrended log-returns on the entire time window 1997-2012. On the y-axis the number of stocks in each cluster is shown, with different colours for different ICB industries. b) For the cluster number 1 in a) we have detected at each time window the correspondent "similar" (according to the hypergeometric test) cluster and we have plotted the composition in time. Size equal to zero corresponds to no "similar" cluster found. When more than one "similar" cluster is found only data of the largest cluster is plotted. c)-f): same plots as in b), for clusters 4, 8, 7 and 17 respectively.

So far we have described the persistence of clusters from a global perspective, looking at the clustering as a whole. Let us here focus on the evolution of each

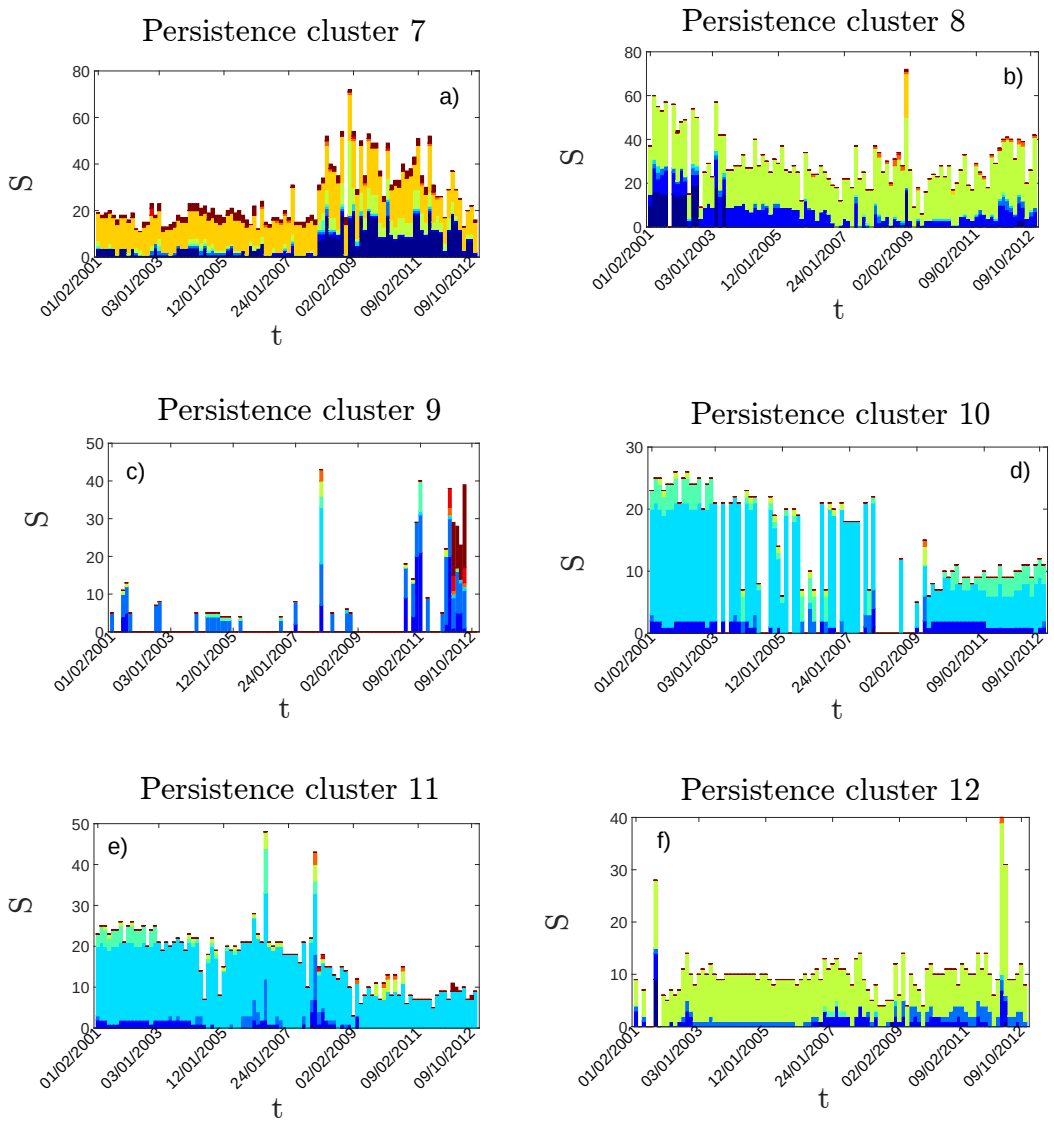

Fig. 5.5 **Clusters dynamical composition (part 2)**. a) For the cluster number 7 in Fig. 5.4 a) we have detected at each time window the correspondent "similar" (according to the hypergeometric test) cluster and we have plotted the composition in time. Size equal to zero corresponds to no "similar" cluster found. When more than one "similar" cluster is found only data of the largest cluster is plotted. b)-f): same plots as in a), for clusters 7, 8, 9, 10, 11 and 12 respectively. Colors refer to the legend in Fig. 5.4 a).

cluster, following how their composition changes in time. It is not straightforward to analyse such an evolution, the main problem being the changeable nature of dynamical clusters that makes difficult to identify the successor for each cluster. Many different approaches can be adopted to address this community tracking problem [195]. Here we use hypothesis statistical tests based on the hypergeometric distribution introduced in

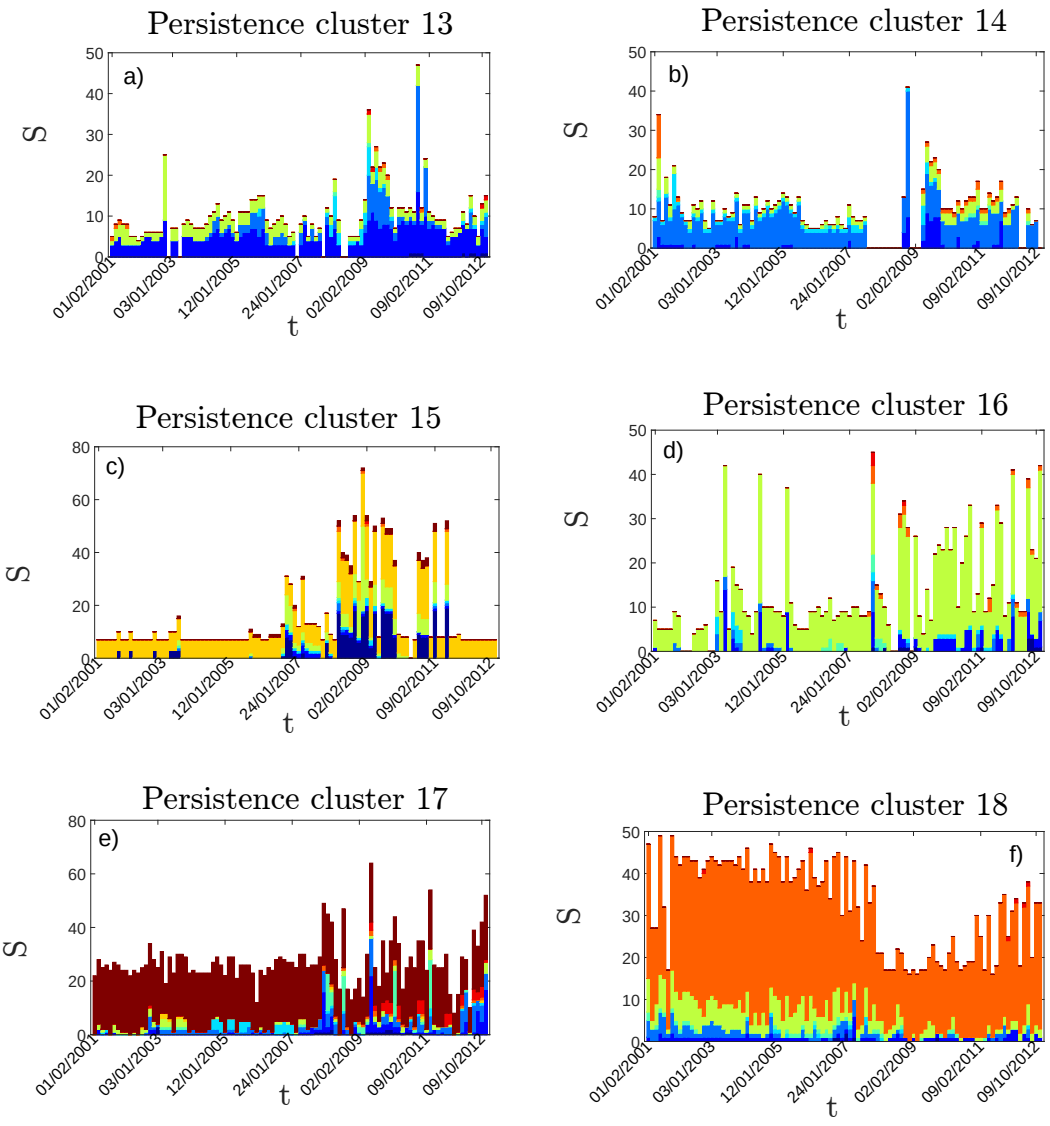

Fig. 5.6 **Clusters dynamical composition (part 3)**. a) For the cluster number 13 in Fig. 5.4 a) we have detected at each time window the correspondent "similar" (according to the hypergeometric test) cluster and we have plotted the composition in time. Size equal to zero corresponds to no "similar" cluster found. When more than one "similar" cluster is found only data of the largest cluster is plotted. b)-f): same plots as in a), for clusters 14, 15, 16, 17 and 18 respectively. Colors refer to the legend in Fig. 5.4 a).

Section 4.2.4 of Chapter 4; while in that chapter we have used this test to find matchings between clusters and ICB supersectors/industries, here the matchings we are interested in are between clusters at different times. In particular, if the number of stocks in common between two clusters is high enough to reject the null hypothesis of the test, we label the two clusters as "similar". Moreover we take the DBHT clustering calculated on

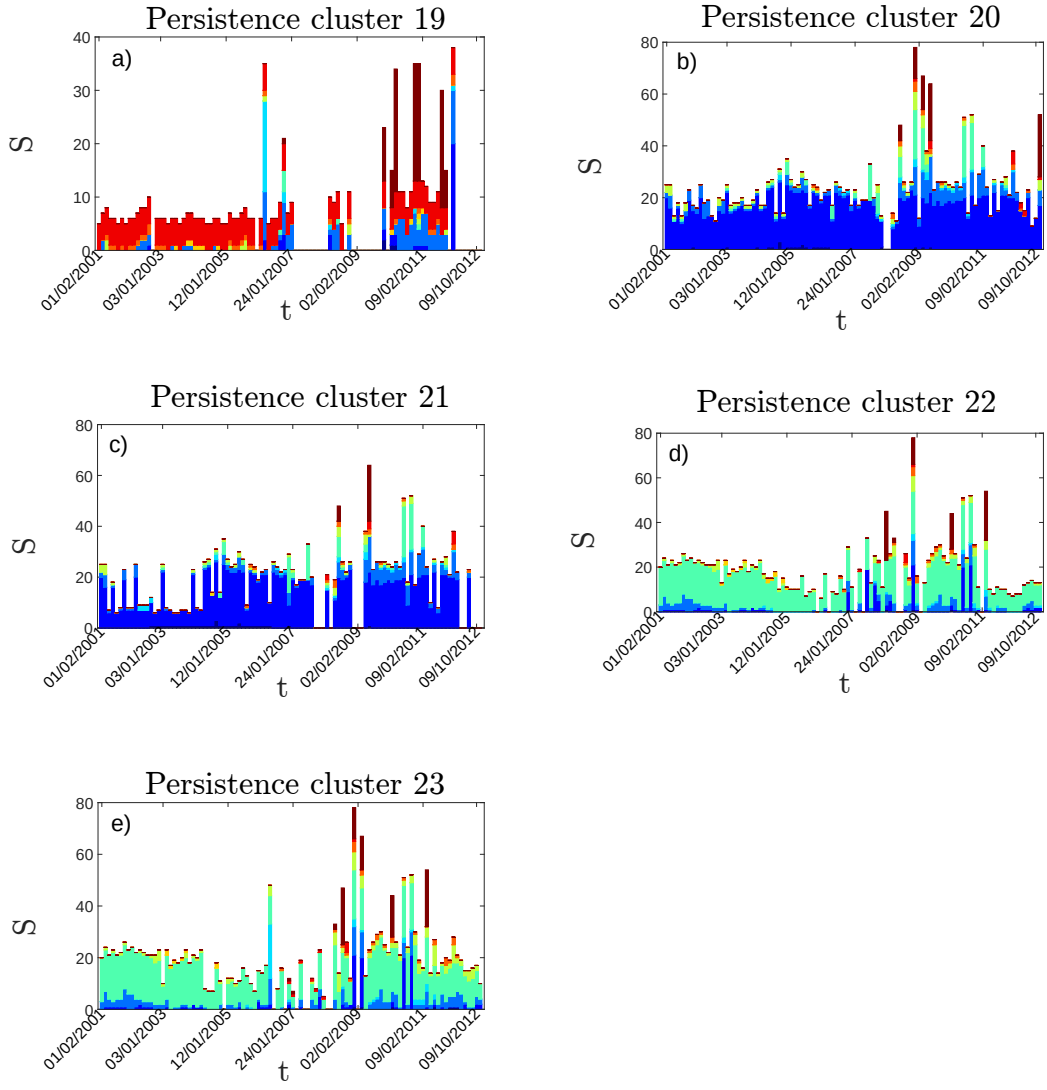

Fig. 5.7 **Clusters dynamical composition (part 4)**. a) For the cluster number 19 in Fig. 5.4 a) we have detected at each time window the correspondent "similar" (according to the hypergeometric test) cluster and we have plotted the composition in time. Size equal to zero corresponds to no "similar" cluster found. When more than one "similar" cluster is found only data of the largest cluster is plotted. b)-f): same plots as in a), for clusters 20, 21, 22 and 23 respectively. Colors refer to the legend in Fig. 5.4 a).

the entire time window (1997-2012) as a benchmark clustering through which tracking the evolution of the dynamical clusters obtained with the moving time windows.

Let us here describe the idea in more details. Let us call X the clustering obtained on the entire time window and Y^i a cluster belonging to X, with $i = 1,...,N_{cl}$. For each cluster Y^i and for each time window T_k ($k = 1,...,n$) we have taken the clustering at

time T_k, X_{T_k}, and identified the cluster belonging to X_{T_k} that is "similar" to Y^i (if any). We label a cluster as "similar" to Y^i if the number of stocks in common with Y^i is high enough to reject the null hypothesis of random overlapping, as quantified by the hypergeometric test [97, 109]. If more than one cluster turns out to be similar to Y^i, we have taken the largest cluster. Eventually we have ended up, for each Y^i, with up to one cluster for each time window T_k, all of them having in common high similarity with Y^i. Through this temporal sequence of clusters we can therefore follow the evolution of Y^i in terms of number of stocks and industrial sector membership. The threshold for each test has been chosen equal to 0.01, together with the conservative Bonferroni correction for multiple tests [9].

In Fig. 5.4 a) the composition of the DBHT clustering X computed on the time window 1997-2012 is shown: for each cluster the y-axis displays its cardinality S, with different colors showing stocks belonging to different ICB industries. In Figs. 5.4 - 5.7 we plot in time the number of stocks S in each cluster, together with their composition in terms of ICB industries. When for a time window no similar clusters can be found we have just left empty the correspondent window. Let us here summarise the main findings:

- There are clusters in X that tend to show quite similar evolutions, for example clusters 5 and 6, or clusters 7 and 15. This means that there are many time windows when these clusters match the same dynamical cluster; in this sense they can be viewed as a single cluster from the dynamical point of view.

- Most of the clusters in X have a high persistence in time, showing a correspondent "similar" dynamical cluster at almost each time window. This result is remarkable as the persistence has been assessed in quite a conservative way, i.e., the hypergeometric test with the Bonferroni correction. Some clusters display a limited number of gaps in their evolution (clusters 14, 15, 19, 20 and 22) in correspondence with the financial crisis. Only clusters 2, 3 and 9 display several gaps not limited to the financial crisis (note that cluster 3 does not show a similar

cluster at any time window, therefore has not been included in the figure). Overall, clusters with several gaps tend to be small clusters, even though not all small clusters have low persistence (e.g. clusters 15 and 16): the graphs suggest that the industrial sector composition play a role as well.

- Few clusters show a persistence in terms of industrial composition as well (it is the case of clusters 4 and, in a less extent, 8), but the majority shows a clear evolution. In particular we can distinguish quite well a pre-crisis and a post-crisis state, the latter characterized by a higher degree of mixing of different industries. If over the pre-crisis period we find clusters dominated by one or two industries (Technology and Industrials in cluster 18, Oil & Gas in 4 and 15, Utilities in 17, Consumer Services and Goods in 14 and 20, Financials in 6, Health Care in 22), in the crisis and post-crisis years the industries tend to mix together much more, forming mixings that were not present earlier (Oil & Gas with Basic Materials and Industrials in cluster 1 and 7, Utilities with Telecommunications and Consumer Services in 17, Financials with Consumer Goods and Services in 6, Health Care with Utilities and Consumer Goods in 20). This again points out the fact that the years after the crisis have seen a drop in the reliability of industries as benchmark to diversify risk.

- Apart from the pre and post-crisis dichotomy, in some cases the 2007-2008 crisis' years show their own features as well. As stated above, some clusters "disappear" during the peak of the crisis (clusters 14, 20 and 22). Many others show instead several peaks in their sizes, together with a sudden increase in the number of industries: this is probably related to the merging of many clusters in few, larger clusters during the crisis.

- The clusters containing Financial stocks (cluster 5 and 6) are worth to analyse further, since they seem to play a role in the outbreak of financial crisis. Indeed they show a clear change in 2007, becoming larger and larger and including an

increasing number of different industries (especially Health Care, Technology and Consumer Services). This pattern is probably connected to the rising importance of the Financial industry as driving factor over the outbreak of crisis. Interestingly at the end of 2008, when Lehman Brothers went bankrupt, this cluster drops suddenly to much lower sizes (although still higher than the pre-crisis values) and less mixed composition. This fact suggests that the Financial industry ends playing a major role in the dependence structure from 2009 onwards.

5.3 Memory in the correlation-based network dynamics

In this section we want to move from the cluster to the PMFG topology, and investigate its temporal evolution. We will focus on metrics drawn from Network Theory, such as degree and clustering coefficient [11, 12]. Since we can measure such metrics for each PMFG at different time windows, we end up with an array of time series that we can analyse with the set of tools available in literature [30]. In particular we are interested in the autocorrelation function, as it is related to the memory properties of the time series and of the dynamical network [30].

In order to have a sufficient number of points for our statistical analysis we have increased the number of time windows by setting the shift dT to 1 trading day and the windows' length to $\theta = 750$ trading days.

The first topological metrics we consider is the degree of each node, that we introduced in Chapter 3. We refer to the degree of asset i at time window T_k with the notation $k_i(T_k)$. Such quantity is highly dynamic for all stocks, as shown in Fig. 5.8 a) for a particular stock (LLTC US Equity). In analogy to the study of financial time series, it is convenient to analyse the variations of such quantity, namely $\Delta k_i(T_k) = k_i(T_k) - k_i(T_{k-1})$. $\Delta k_i(T_k)$ is shown in Fig. 5.8 b) for LLTC US Equity. In order to study the memory

properties of the degree evolution we compute the sample autocorrelation function for $\Delta k_i(T_k)$ as follows [30]:

$$r(\tau) = \frac{\sum_{k=1}^{n-\tau}(\Delta k_i(T_k) - \overline{\Delta k_i})(\Delta k_i(T_{k+\tau}) - \overline{\Delta k_i})}{\sum_{k=1}^{n}(\Delta k_i(T_k) - \overline{\Delta k_i})^2} \tag{5.3}$$

where τ is the lag, n the total number of time windows and $\overline{\Delta k_i}$ is the sample mean of degree variations:

$$\overline{\Delta k_i} = \frac{1}{n}\sum_{k=1}^{n}\Delta k_i(T_k) \tag{5.4}$$

In Fig. 5.8 c) $r(\tau)$ is shown. As we can see, at $\tau = 1$ we find a significant anticorrelation; this indicates a mean reversion property in the evolution of node's degree [30]. For $\tau > 1$ there are no significant correlations. However we find a richer structure if we turn to the autocorrelation of absolute values of $\Delta k_i(T_k)$:

$$r^{abs}(\tau) = \frac{\sum_{k=1}^{n-\tau}(|\Delta k_i(T_k)| - \overline{|\Delta k_i|})(|\Delta k_i(T_{k+\tau})| - \overline{|\Delta k_i|})}{\sum_{k=1}^{n}(|\Delta k_i(T_k)| - \overline{|\Delta k_i|})^2} \quad , \tag{5.5}$$

where now $\overline{|\Delta k_i|}$ is the sample mean of absolute value of degree variations:

$$\overline{|\Delta k_i|} = \frac{1}{n}\sum_{k=1}^{n}|\Delta k_i(T_k)| \quad . \tag{5.6}$$

In Fig. 5.8 d) this autocorrelation is shown and we can observe a significant - although weak - correlation that spans more than one month. This memory structure resembles the volatility clustering of asset returns discussed in Section 2.3 of Chapter 2: loosely speaking, high variations of degree are likely to be followed by high variations, and low variations are likely to be followed by low variations. Moreover, all these properties are not unique of the degree. We have carried out the same study by using a more complex topological metrics, namely the clustering coefficient of each node, $c_i(T_k)$: qualitatively the same properties are found, as shown in Fig. 5.9 a)-d).

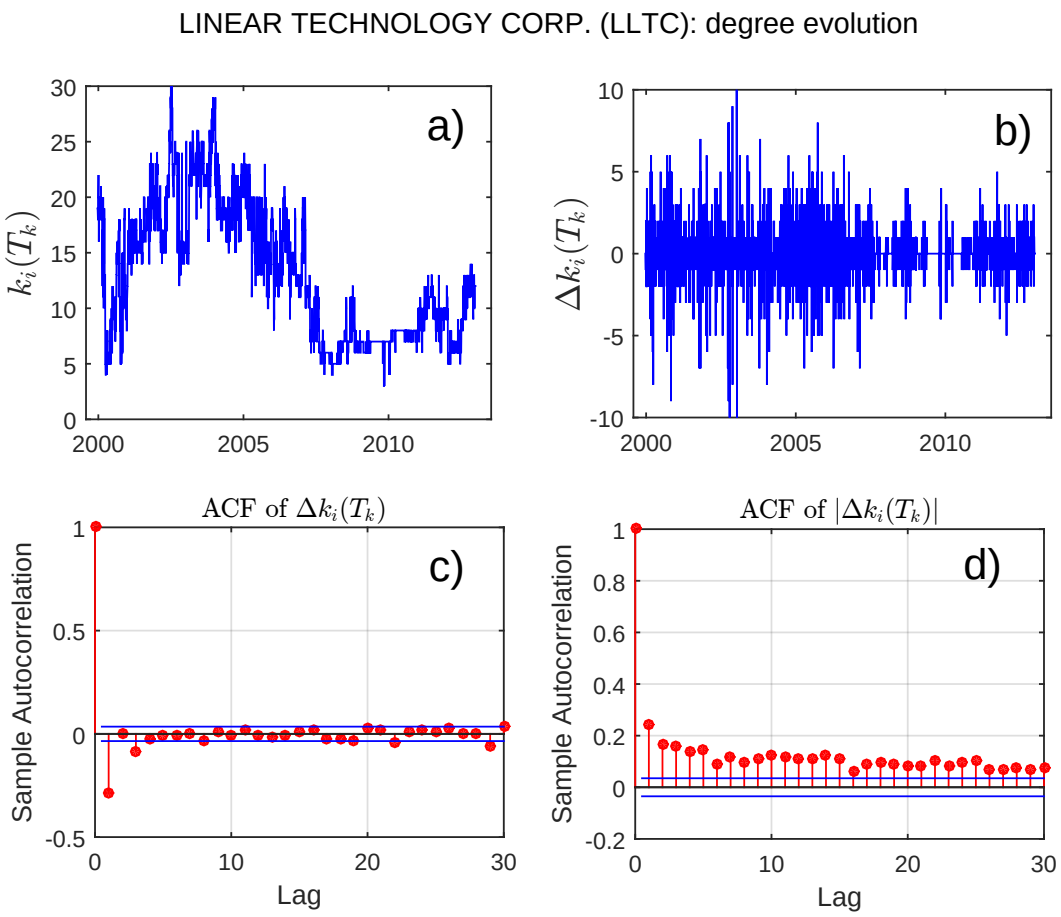

Fig. 5.8 **Analysis of degree evolution for LLTC US Equity**. a) Degree $k_i(T_k)$ as a function of time window T_k. b) Degree variation $\Delta k_i(T_k)$ as a function of time window T_k; a volatility clustering effect is evident. c) Autocorrelation function (ACF) of degree variation $\Delta k_i(T_k)$; all lags show no autocorrelation but lag 1, where a negative autocorrelation is observed. d) Autocorrelation function (ACF) of absolute value of degree variation $|\Delta k_i(T_k)|$; here significant autocorrelation can be found across a wide range of lags, confirming the volatility clustering property.

In order to investigate further the autocorrelation of absolute variations we have analysed the functional form of its decay. For the asset returns, the autocorrelation decay of absolute values is known to be roughly power law [5]. For what concerns the degree and clustering coefficient, we do not find strong evidence of power law decay for all assets: the average R^2 for a power law fit is about 0.68 for the degree and 0.7 for the clustering coefficient (see Figs. 5.10 a) and b)). However, around 67% of stocks for the degree and 91% of stocks for the clustering coefficient provide a better regression with a power law than an exponential decay, as shown in Figs. 5.10 c) and d). In terms

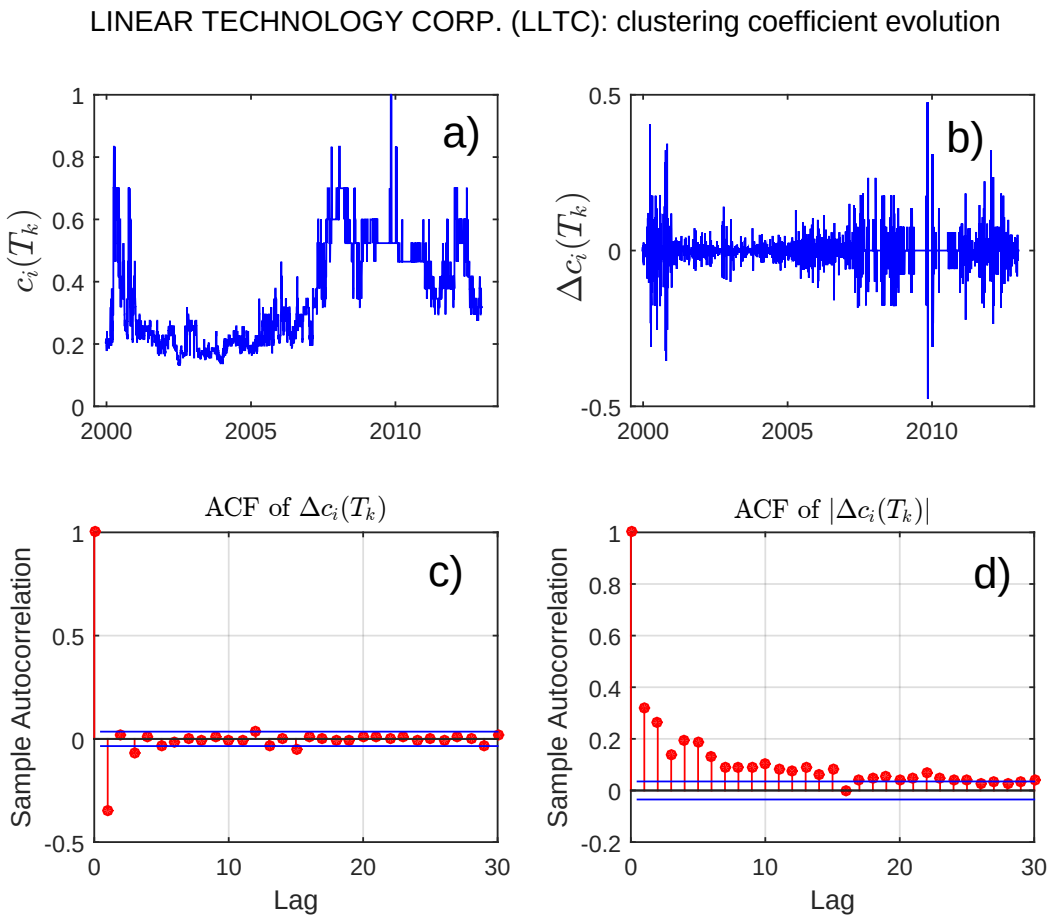

Fig. 5.9 **Analysis of clustering coefficient evolution for LLTC US Equity**. a) Clustering coefficient $c_i(T_k)$ as a function of time window T_k. b) Clustering variation $\Delta c_i(T_k)$ as a function of time window T_k; a volatility clustering effect is evident. c) Autocorrelation function (ACF) of clustering coefficient variation $\Delta c_i(T_k)$; all lags show no autocorrelation but lag 1, where a negative autocorrelation is observed. d) Autocorrelation function (ACF) of absolute value of clustering coefficient variation $|\Delta c_i(T_k)|$; here significant autocorrelation can be found across a wide range of lags, confirming the volatility clustering property.

of power law exponents α, both distributions are quite wide, ranging from -2.33 to -0.147 (see Figs. 5.10 e) and f)); the mean α is -1.58 for the degree and -1.42 for the clustering coefficient.

These analyses indicate the presence of long-term memory in the evolution of the dependence structure; such feature could be the foundation for a first attempt to model the evolution of correlation-based networks.

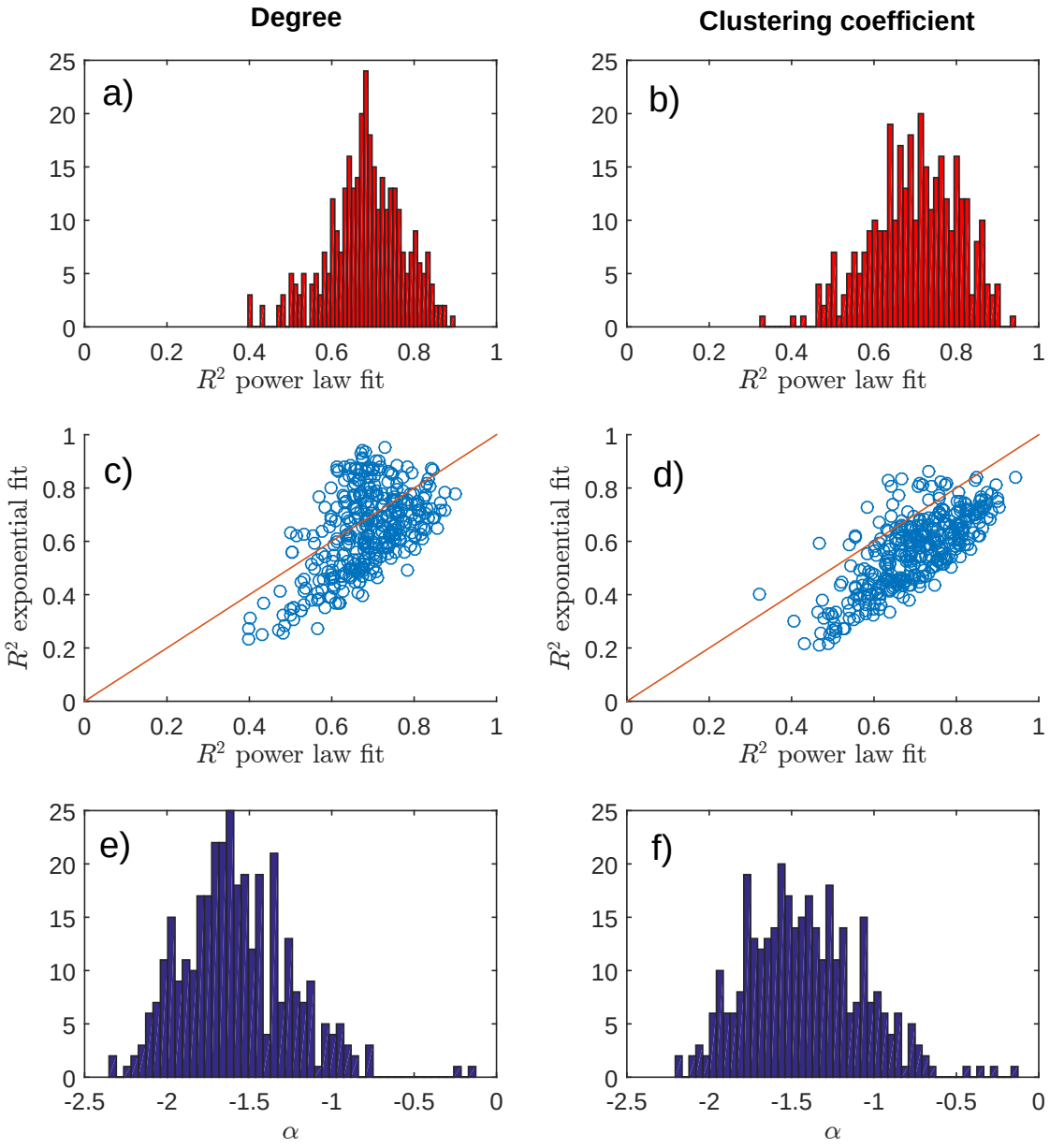

Fig. 5.10 **Summary of regression analysis for the autocorrelation functions decay**. a)-b): Histograms of R^2 obtained from power-law regression of autocorrelation functions of $|\Delta k_i(T_k)|$ a) and $|\Delta c_i(T_k)|$ b). c)-d): Comparison between R^2 obtained from power-law (x-axis) and exponential (y-axis) decay, for $|\Delta k_i(T_k)|$ and $|\Delta c_i(T_k)|$ respectively; the straight line represents the bisecting line $y = x$; around 67% and 91% of stocks provide a higher R^2 for power-law than exponential regression. e)-f): Histograms of power-law exponent α for $|\Delta k_i(T_k)|$ and $|\Delta c_i(T_k)|$ respectively.

5.4 Summary

In this chapter we have investigated the dynamical evolution and non-stationarity of market dependence structure by means of correlation-based filtered networks. In particular, we have focused on PMFG and the clustering that its topology naturally provides by means of the Directed Bubble Hierarchical Tree (DBHT) method. We have first focused on the clustering; in particular we have measured the persistence of the dependence structure by calculating similarity among clusterings at different time windows, using the Adjusted Rand Index for quantifying the similarity. On a more refined level, we have tracked the evolution of each single cluster by using the hypergeometric hypothesis test. We have then investigated the dynamic evolution of the PMFGs, through a time series analysis on network metrics such as degree and clustering coefficient.

The analyses reveal that the outbreak of the 2007-2008 financial crisis marks a transition from relatively high levels of persistence to a much more unstable and changeable structure. We have found that the minimum persistence is reached at the end of 2008 when the crisis was fully unfolded. But the decay in persistence started already in the late 2006 well before other warning signs were detectable. The dependence structure persistence eventually recovered in the second half of 2009 with relatively high values until the end of 2011. However, we have shown that such a persistent structure had distinct features from the pre-crisis structure, in particular lower relations with the industrial sectors activities. Notably, since the end of 2011 we are observing a new decay in persistence which is signaling the building-up of another unfolding change in the market structure. This also points out that from 2007 onwards correlation matrices from historical data, both filtered and unfiltered, have become more unstable and therefore less reliable instruments for risk diversification. Furthermore, our analysis on the evolving industrial sector composition of each single cluster reveals that most of them display a clear change with the crisis, that overall makes them more heterogeneous in terms of industrial sectors. In particular, we observed that one cluster, mainly made

of Financial stocks, experiences a sharp rise in its size and heterogeneity that is probably a picture of the breakdown of late 2007 financial crisis. This could give interesting insights in terms of early warning signals.

Finally, with the analysis on the PMFGs evolution we have shown that the dependence structure dynamics is characterised by long term memory. In particular we have found evidence of long range autocorrelation in the absolute values of degree and clustering coefficient variation, reflecting a phenomenon analogous to the volatility clustering in log-returns. This finding opens interesting scenarios for the modeling of correlation-based networks dynamics, as well as for the forecasting of financial correlation. In the next chapter we investigate further how the correlation-based networks evolution can give insight into the future structure of dependence.

Chapter 6

A new approach to volatility forecasting

In this chapter we show how correlation-based networks can be used to forecast changes in the market volatility. We introduce a new measure, the "dependence structure persistence", which is based on the PMFG and which turns out to be a good predictor for volatility variations. We discuss the possible motivation behind such connection, and assess the goodness of this forecasting tool by means of out-of-sample tests on two different data sets. This result is the first step towards the application of correlation-based networks to modeling and forecasting, beyond the descriptive analyses they have been used for so far. The results and analyses presented in this chapter are based on the paper "What does past correlation structure tell us about the future? An answer from network filtering", that has been submitted to a peer-reviewed scientific journal in 2016.

6.1 Introduction

Models for describing and forecasting the evolution of the volatility and covariance among financial assets are widely applied in industry [113, 46, 6]. Among the most popular approaches are worth mentioning the multivariate extensions of GARCH [44], the stochastic covariance models [45] and realized covariance [196]. However most of

these econometrics tools are not able to cope with more than few assets, due to the curse of dimensionality and increase in the number of parameters [113]. Their insight into the volatility evolution is therefore limited to baskets of few assets: they fail to describe or predict the volatility of a portfolio made of hundreds of assets. This is unfortunate, since modeling the evolution of entire markets would provide valuable insights into systemic risk and the unfolding of financial crises [113].

We suggest that the network filtering can be a valuable tool to overcome this limitation. Indeed, the volatility of a portfolio depends on the covariance matrix of the corresponding assets [111]. Therefore, by tracking the evolution of the dependence structure, correlation-based networks can provide insights into future values of volatility. Yet, so far the network filtering has been used mostly for descriptive analyses, with the connections with risk forecasting being mostly overlooked. Some works have shown that is possible to use dimensionality reduction techniques, such as spectral methods applied on correlation matrices, as early-warning signals for systemic risk [197, 198]: however these approaches, although promising, do not provide proper forecasting tools, as they are affected by high false positive ratios and are not designed to predict a specific quantity.

In this chapter we propose an approach which exploits network filtering to explicitly predict future volatility of markets made of hundreds of stocks. To this end, we introduce a new dynamic measure that quantifies the rate of change in the structure of the market correlation matrix: the dependence structure persistence $\langle ES \rangle$. This quantity is derived from past correlation data after that network filtering is performed. Then we show how such measure exhibits significant predicting power on the market volatility, providing a tool to forecast it. We assess the reliability of this forecasting through out-of-sample tests on two different data sets of equity data.

The original contributions of this chapter are the following:

- We introduce the "dependence structure persistence", a new measure based on network-filtering measures, which quantifies the rate of change of the dependence

structure. This measure is a valuable tool for detecting structural changes in the market which could affect the risk evaluation.

- We demonstrate that this index provides information on future variations of market volatility; we assess this relation through a block-bootstrapping analysis [199].

- We propose a method which exploits this relation to forecast future market volatility by using past correlation. We assess the predicting power of this forecasting method through an out-of-sample analysis.

The rest of this chapter is structured as follows: in Section 6.2 we describe the two data sets we have used for the analyses; in Section 6.3 we introduce the "dependence structure persistence"; in Section 6.4 we discuss how this index is related to variation in market volatility, and we assess the significance of this relation through a block-bootstrapping analysis [199]; in Section 6.5 we describe how this relation can be exploited to provide a forecasting tool useful for risk management, by presenting out-of-sample tests [14] and false positive analyses [200].

6.2 Data sets: US and UK data

We have analysed two different data sets of equity data. The first set (NYSE data set) is the data set we have introduced in Chapter 2, composed by daily prices of $N = 342$ US stocks traded in the New York Stock Exchange from 02/01/1997 to 31/12/2012. The second set (LSE dataset) is analysed in this chapter for the first time in this thesis: it is composed by daily prices of $N = 214$ UK stocks traded in the London Stock Exchange, covering 13 years from 05/01/2000 to 21/08/2013. All stocks have been continuously traded throughout this period of time.

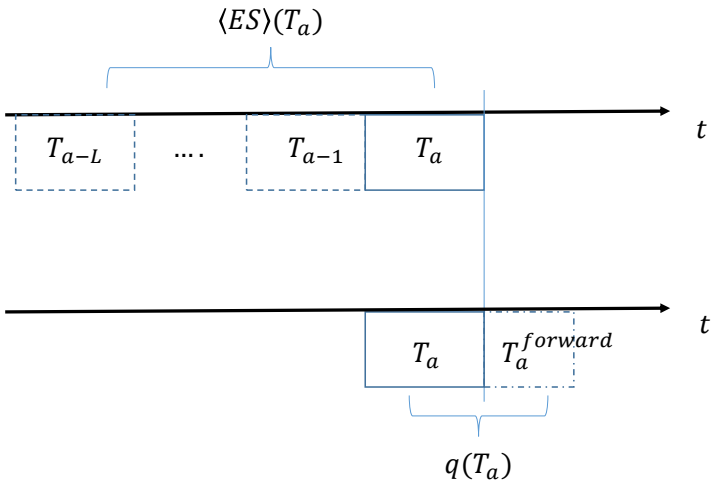

Fig. 6.1 **Scheme of time windows setting**. In the backward-looking setting (upper axis) the time windows are actually overlapping, but they are here represented as disjoint for the sake of simplicity.

6.3 A measure of dependence structure persistence

Let us assume we have computed a dynamic set of correlation matrices $\{\rho_{ij}(T_k)\}$, from the rolling time window set-up described in Section 2.4.4. From each correlation matrix we then compute the corresponding PMFG [55]. Once the n PMFGs, $G(T_k)$ with $k = 1, ..., n$, have been computed we calculate a measure that monitors the dependence structure persistence, based on a measure of PMFG similarity. Such measure is defined as follows:

$$\langle ES \rangle(T_k) = \sum_{k'=k-L}^{k-1} \omega(T_{k'}) ES(T_k, T_{k'}) \ , \tag{6.1}$$

where $\omega(T_{k'}) = \omega_0 \exp(\frac{k'-k-1}{L/3})$ is an exponential smoothing factor, L is a parameter and $ES(T_k, T_{k'})$ is the fraction of edges in common between the two PMFGs $G(T_k)$ and $G(T_{k'})$, called "edge survival ratio" [76]. In formula, $ES(T_k, T_{k'})$ reads:

$$ES(T_k, T_{k'}) = \frac{1}{N_{edges}} \mid E^{T_k} \cap E^{T_{k'}} \mid, \tag{6.2}$$

where N_{edges} is the number of edges (links) in the two PMFGs (constant and equal to $3N-6$ for a PMFG [55]), and E^{T_k} ($E^{T_{k'}}$) represents the set of edges of PMFG at T_k ($T_{k'}$). The dependence structure persistence $\langle ES \rangle (T_k)$ is therefore a weighted average of the similarity (as measured by the edge survival ratio) between $G(T_k)$ and the first L previous PMFGs, with an exponential smoothing scheme that gives more weight to those PMFGs that are closer to T_k. The parameter ω_0 in Eq. 6.1 can be calculated by imposing $\sum_{k'=k-L}^{k-1} \omega(T_{k'}) = 1$. Intuitevely, $\langle ES \rangle (T_k)$ measures how slowly the change of the dependence structure is occuring in the near past of T_k.

6.4 Dependence analysis

To investigate the relation between $\langle ES \rangle (T_k)$ and the market volatility evolution, let us here introduce another quantity, namely the volatility ratio $q(T_k)$ [147]. Because of non-stationarity, the covariance estimated from historical data on a time window T_k is not always a good proxy for the covariance measured in a future time window $T_k^{forward}$ (the so-called realized covariance). This translates into an uncertainty on the risk in the market, as measured by the volatility [110]. In order to quantify the agreement between the estimated and the realized risk we here make use of the volatility ratio, a measure which has been used in [147, 201, 38] for this purpose and defined as follows:

$$q(T_k) = \frac{\sigma(T_k^{forward})}{\sigma(T_k)}, \tag{6.3}$$

where $\sigma(T_k^{forward})$ is the realized volatility of the average market return $r_M(t)$ computed on the time window $T_k^{forward}$; $\sigma(T_k)$ is the estimated volatility of $r_M(t)$ computed on time window T_k, by using the same exponential smoothing scheme [103] described for the correlation $\{\rho_{ij}(T_k)\}$. Specifically, $T_k^{forward}$ is the time window of length $\theta_{forward}$ that follows immediately T_k: if t_θ is the last observation in T_k, $T_k^{forward}$ covers observations from $t_{\theta+1}$ to $t_{\theta+1+\theta_{forward}}$ (Fig. 6.1). Therefore the ratio in Eq. 6.3 estimates the agreement between the market volatility estimated with observations in

T_k and the actual market volatility observed over an investment in the N assets over $T_k^{forward}$. If $q(T_k) > 1$, then the historical data gathered at T_k has underestimated the (future) realized volatilty, whereas $q(T_k) < 1$ indicates overestimation.

Let us stress that $q(T_k)$ provides an information on the reliability of the covariance estimation too, given the relation between market return volatility and covariance [111]:

$$\sigma(T_k) = \sqrt{\frac{1}{N^2}\sum_{ij}\Sigma_{ij}(T_k)}, \tag{6.4}$$

$$\sigma(T_k^{forward}) = \sqrt{\frac{1}{N^2}\sum_{ij}\Sigma_{ij}(T_k^{forward})}, \tag{6.5}$$

where $\Sigma_{ij}(T_k)$ and $\Sigma_{ij}(T_k^{forward})$ are respectively the estimated and realized covariances.

To investigate the relation between $\langle ES \rangle(T_k)$ and $q(T_k)$ we have calculated the two quantities with different values of θ and L in Eqs. 6.1 and 6.3, to assess the robustness against these parameters. Specifically, we have used $\theta \in (250, 500, 750, 1000)$ trading days, that correspond to time windows of length 1, 2, 3 and 4 years respectively; $L \in (10, 25, 50, 100)$, that correspond (given $dT = 5$ trading days) to an average in Eq. 6.1 reaching back to 50, 125, 250 and 500 trading days respectively.

In Fig. 6.2 we show the $ES(T_k, T_{k'})$ matrices (Eq. 6.2) for the NYSE and LSE datasets, for $\theta = 1000$. We can observe a block structure much similar to what observed in Chapter 5 by using the Adjusted Rand Index. Similar structures are found for all values of θ considered. In Fig. 6.3 we show $\langle ES \rangle(T_k)$ and $q(T_k)$ as a function of time, for $\theta = 1000$ and $L = 100$. As expected, main peaks of $q(T_k)$ occur during the months before the most turbulent periods in the stock market (the 2002 market downturn and the 2007-08 credit crisis), when the underestimation of future volatility is very high. Interestingly, $\langle ES \rangle(T_k)$ seems to follow a specular trend. This is confirmed by explicit calculation of Pearson correlation between the two signals, reported in Tabs. 6.1 - 6.2: as one can see, for all combinations of parameters the correlation is negative.

6.4.1 Test of significance: block-bootstrapping

In order to check the significance of this anticorrelation we cannot rely on standard tests on Pearson coefficient, such as Fisher transform [202], as they assume i.i.d. series [138]. Our time series are instead strongly autocorrelated, due to the overlapping between adjacent time windows. Therefore we have calculated confidence intervals by performing a block bootstrapping test [199]. This is a variation of the bootstrapping test [190], conceived to take into account the autocorrelation structure of the original series. The only free parameter in this method is the block length, that we have chosen applying the optimal selection criterion proposed in [203]: such criterion is adaptive on the autocorrelation strength of the series as measured by the correlogram. In our case we have found, depending on the parameters θ and L, optimal block lengths ranging from 29 to 37, with a mean of 34 (corresponding to 170 trading days). By performing block bootstrapping tests we have therefore estimated confidence intervals for the true correlation between $\langle ES \rangle(T_k)$ and $q(T_k)$; in Tabs. 6.1 - 6.2 correlations whose 95% and 99% confidence intervals (CI) do not include zero are marked with one and two stars respectively. As we can see, 14 out of 16 correlation coefficients are significantly different from zero within 95% CI in the NYSE dataset, and 12 out of 16 in the LSE dataset. For what concerns the 99% CI, we observe 13 out 16 for the NYSE and 9 out of 16 for the LSE dataset. Non-significant correlations appear only for $\theta = 250$, suggesting that this length is too small to provide a reliable measure of structural persistence. Very similar results are obtained by using Minimum Spanning Tree (MST) [167] instead of PMFG as correlation-based filtered network.

Given the interpretation of $\langle ES \rangle(T_k)$ and $q(T_k)$ given above, anticorrelation implies that an increase in the "speed" of dependence structure evolution (low $\langle ES \rangle(T_k)$) is likely to correspond to underestimation of future market volatility from historical data (high $q(T_k)$), whereas when the structure evolution "slows down" (high $\langle ES \rangle(T_k)$) there is indication that historical data is likely to provide an overestimation of future volatility. This means that we can use $\langle ES \rangle(T_k)$ as a valuable predictor of current historical data

reliability. This result is to some extent surprising as $\langle ES \rangle (T_k)$ is derived from PMFGs topology, that in turns depends only on the ranking of correlations and not on their actual value: yet, this information provides meaningful information about the future market volatility and therefore about the future covariance.

6.4.2 The advantage of network filtering

In principle other measures of correlation ranking structure, more straightforward than the correlation persistence $\langle ES \rangle (T_k)$, might capture the same interplay with $q(T_k)$. We here considered the Metacorrelation $z(T_k, T_{k'})$, that is the Pearson correlation computed between the coefficients of correlation matrices at T_k and $T_{k'}$ [153]. This measure does not make use of PMFG and is defined as follows:

$$z(T_k, T_{k'}) = \frac{\langle \rho_{ij}(T_k) \rho_{ij}(T_{k'}) \rangle_{ij}}{\sqrt{[\langle \rho_{ij}^2(T_k) \rangle_{ij} - \langle \rho_{ij}(T_k) \rangle_{ij}^2][\langle \rho_{ij}^2(T_{k'}) \rangle_{ij} - \langle \rho_{ij}(T_{k'}) \rangle_{ij}^2]}}, \tag{6.6}$$

where $\rho_{ij}(T_k)$ is the correlation between stocks i and j at time window T_k and $\langle ... \rangle_{ij}$ is the average over all pairs of stocks i, j. Fig. 6.4 displays the similarity matrices obtained with this measure for NYSE and LSE datasets: we can observe again block-like structures, that however carry different information from the $ES(T_k, T_{k'})$ in Fig. 6.2; in particular, blocks show higher intra-similarity and less structure.

Similarly to Eq. 6.1 we have then defined $z(T_k)$ as the weighted average over L past time windows:

$$\langle z \rangle (T_k) = \sum_{b=a-L}^{a-1} \omega(T_{k'}) z(T_k, T_{k'}). \tag{6.7}$$

In Tabs. 6.3 and 6.4 we show the correlation between $z(T_k)$ and $q(T_k)$. As we can see, although an anticorrelation is present for each combination of parameters θ and L, correlation coefficients are systematically closer to zero than in Tabs. 6.1 - 6.2, where correlation persistence was used. Moreover the number of significant Pearson coefficients, according to the block bootstrapping, decreases to 12 out of 16 in NYSE

and to 10 out of 16 in LSE dataset. Since $\langle z \rangle(T_k)$ does not make use of PMFG, this result suggests that the filtering procedure associated to correlation-based networks is a necessary step for capturing at best the correlation ranking evolution and its interplay with the volatility ratio.

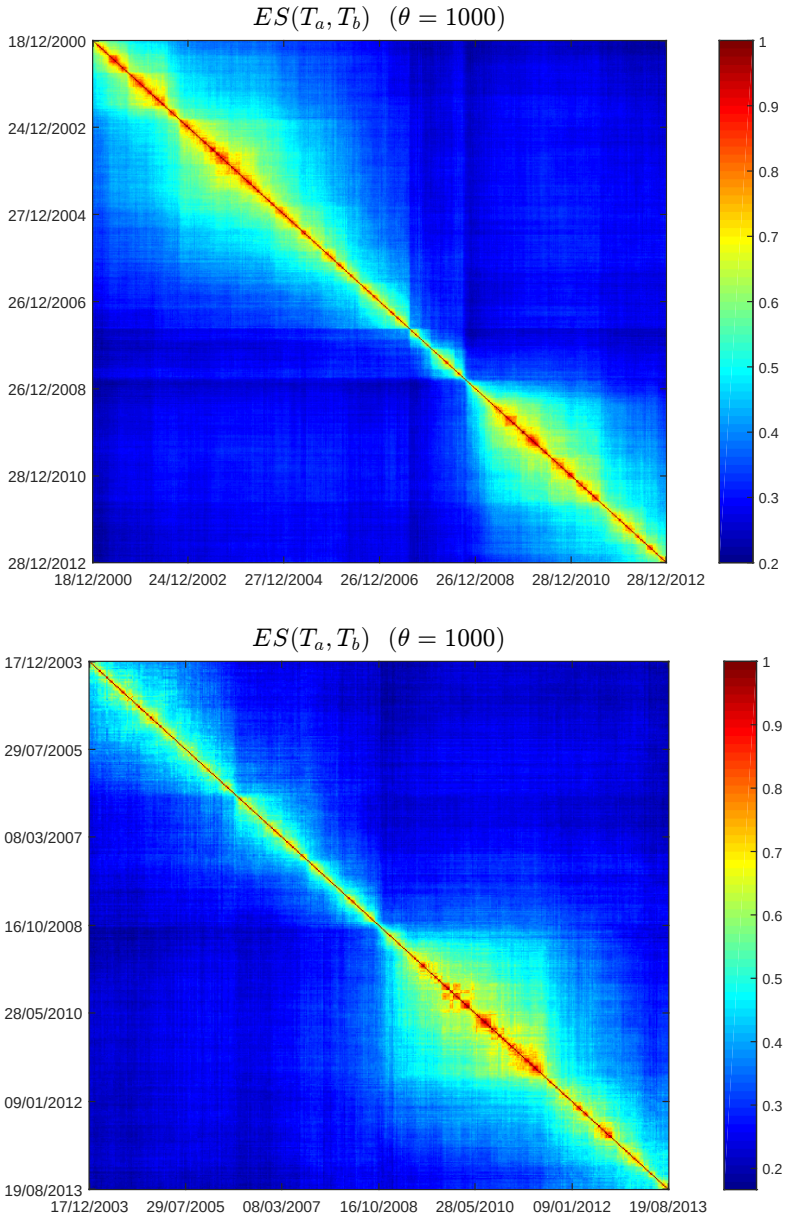

Fig. 6.2 $ES(T_k, T_{k'})$ **matrices for** $\theta = 1000$**, for NYSE (left) and LSE dataset (right)**. A block-like structure can be observed in both datasets, with periods of high structural persistence and other periods whose dependence structure is changing faster. The 2007-2008 financial crisis marks a transition between two main blocks of high structural persistence.

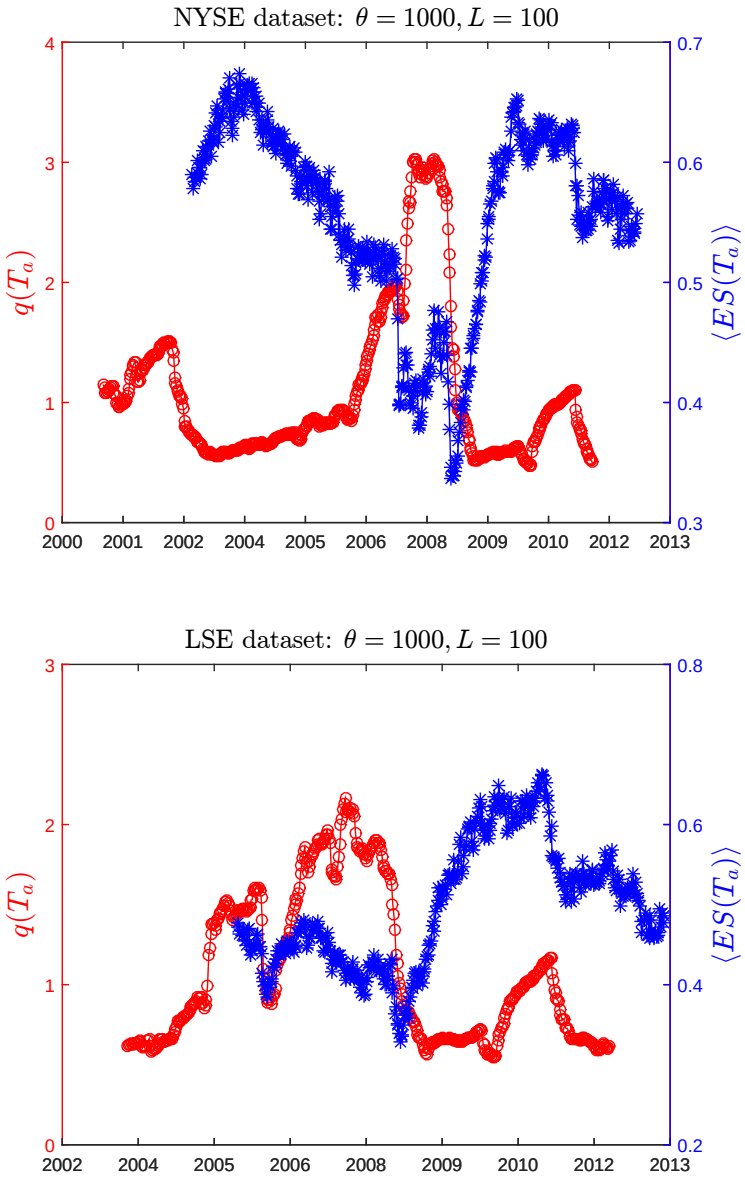

Fig. 6.3 $\langle ES\rangle(T_k)$ **and** $q(T_k)$ **signals represented for** $\theta = 1000$ **and** $L = 100$, for both NYSE (left graph) and LSE (right graph) datasets. It is evident the anticorrelation between the two signals. The financial crisis triggers a major drop in the structural persistence and a corresponding peak in $q(T_k)$.

6.5 Forecasting

In this section we evaluate how well the dependence structure persistence $\langle ES\rangle(T_k)$ can forecast the future through its relation with the forward-looking volatility ratio $q(T_k)$. In particular we focus on estimating whether $q(T_k)$ is greater or less than 1:

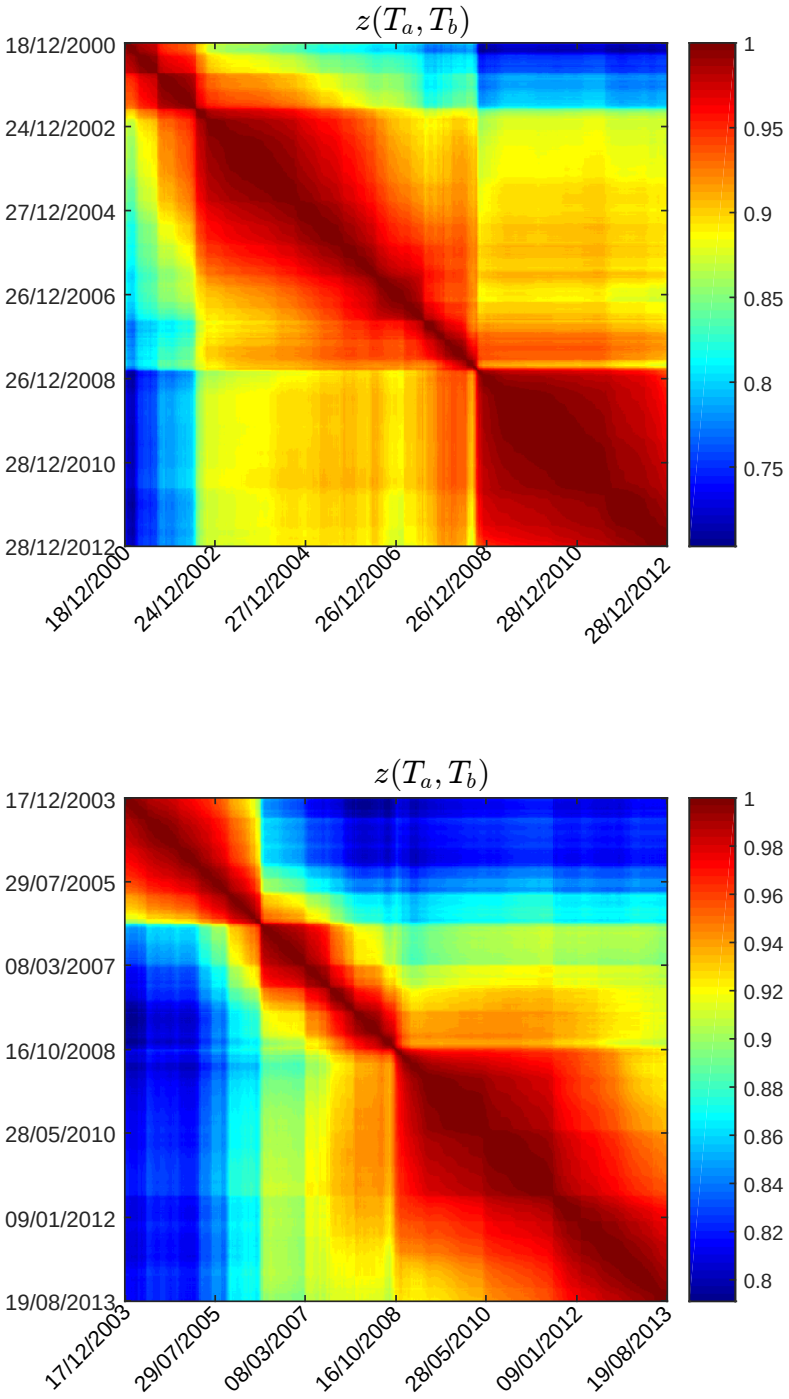

Fig. 6.4 $z(T_k, T_{k'})$ **matrices for** $\theta = 1000$**, for NYSE (left) and LSE dataset (right).** A block-like structure can be observed in both datasets, with periods of high structural persistence and other periods whose dependence structure is changing faster. The blocks of high similarity show higher compactness than in Fig. 6.2.

Table 6.1 **NYSE dataset: correlation between $\langle ES\rangle(T_a)$ and $q(T_a)$**, for different combinations of parameters θ and L. Stars mark those correlation coefficients whose confidence interval excludes zero with a 95% (one star) or a 99% confidence (two stars). The confidence intervals are computed from the block-bootstrapped sample.

		L			
		10	**25**	**50**	**100**
θ	**250**	-0.2129	-0.2224	-0.2997^{*}	-0.3498^{**}
	500	-0.4276^{**}	-0.4683^{**}	-0.4945^{**}	-0.5354^{**}
	750	-0.4994^{**}	-0.5499^{**}	-0.5837^{**}	-0.6018^{**}
	1000	-0.5789^{**}	-0.6152^{**}	-0.6480^{**}	-0.6874^{**}

$^{**}p < 0.001$, $^{*}p < 0.01$,

Table 6.2 **LSE dataset: correlation between $\langle ES\rangle(T_a)$ and $q(T_a)$**, for different combinations of parameters θ and L. Stars mark those correlation coefficients whose confidence interval excludes zero with a 95% (one star) or a 99% confidence (two stars). The confidence intervals are computed from the block-bootstrapped sample.

		L			
		10	**25**	**50**	**100**
θ	**250**	-0.2084^{*}	-0.1887^{*}	-0.1872	-0.2269^{*}
	500	-0.3083^{**}	-0.3343^{**}	-0.3782^{**}	-0.4202^{**}
	750	-0.4050^{**}	-0.4409^{**}	-0.4334^{**}	-0.4374^{**}
	1000	-0.4552^{**}	-0.5285^{**}	-0.5480^{**}	-0.5227^{**}

$^{**}p < 0.001$, $^{*}p < 0.01$,

this information, although less complete than a precise estimation of $q(T_k)$, gives us an important insight into possible overestimations ($q(T_k) < 1$) or underestimation ($q(T_k) > 1$) of future volatility.

We proceed as follows. Given a choice of parameters θ and L, we calculate the corresponding set of pairs $\{\langle ES\rangle(T_k), q(T_k)\}$, with $k = 1, ..., n$. Then we define $Y(T_k)$ as the categorical variable that is 0 if $q(T_k) < 1$ and 1 if $q(T_k) > 1$. Finally we perform a logistic regression of $Y(T_k)$ against $\langle ES\rangle(T_k)$: namely, we assume that [14]:

$$P\{Y(T_k) = 1 | \langle ES\rangle(T_k) = x\} = S(\beta_0 + \beta_1 x) \quad , \tag{6.8}$$

where $S(t)$ is the sigmoid function $S(t) = \frac{1}{1+e^{-t}}$ [204]; we estimate parameters β_0 and β_1 from the observations $\{\langle ES\rangle(T_k), q(T_k)\}_{k=1,...,n}$ through Maximum Likelihood [15].

Table 6.3 **NYSE dataset: correlation between $\langle z \rangle(T_a)$ and $q(T_a)$**, for different combinations of parameters θ and L. Stars mark those correlation coefficients whose confidence interval excludes zero with a 95% (one star) or a 99% confidence (two stars). The confidence intervals are computed from the block-bootstrapped sample.

		L			
		10	**25**	**50**	**100**
θ	**250**	-0.0992	-0.0754	-0.1055	-0.1157
	500	-0.2146	-0.2232	-0.2309	-0.2753
	750	-0.2997	-0.3706^{*}	-0.4030^{*}	-0.4109^{*}
	1000	-0.3933^{**}	-0.4290^{**}	-0.4678^{**}	-0.4574^{*}

$^{**}p < 0.001$, $^{*}p < 0.01$,

Table 6.4 **LSE dataset: correlation between $\langle z \rangle(T_a)$ and $q(T_a)$**, for different combinations of parameters θ and L. Stars mark those correlation coefficients whose confidence interval excludes zero with a 95% (one star) or a 99% confidence (two stars). The confidence intervals are computed from the block-bootstrapped sample.

		L			
		10	**25**	**50**	**100**
θ	**250**	-0.1470	-0.1095	-0.1326	-0.1720
	500	-0.2365^{*}	-0.2113	-0.2936^{*}	-0.3932^{**}
	750	-0.3123^{**}	-0.3379^{*}	-0.3538^{*}	-0.3851^{*}
	1000	-0.2917^{*}	-0.2954	-0.3163	-0.4192^{**}

$^{**}p < 0.001$, $^{*}p < 0.01$,

Once the model is calibrated, given a new observation $\langle ES \rangle(T_{n+1}) = x$ we predict $Y(T_{n+1}) = 1$ if $P\{Y(T_{n+1}) = 1 | \langle ES \rangle(T_{n+1}) = x\} > 0.5$, and $Y(T_{n+1}) = 0$ otherwise. This classification criterion, in a case with only one predictor, corresponds to classify $Y(T_{n+1})$ according to whether $\langle ES \rangle(T_{n+1})$ is greater or less than a threshold r which depends on β_0 and β_1, as shown in Figs. 6.5-6.6 for a particular choice of parameters. Therefore the problem of predicting whether market volatility will increase or decrease boils down to a classification problem [15] with $\langle ES \rangle(T_k)$ as predictor and $Y(T_k)$ as target variable.

We make use of a logistic regression because it is more suitable than a polynomial model for dealing with classification problems [14]. Other classification algorithms are available; we have chosen the logistic regression due to its simplicity. We have also implemented the KNN algorithm [15] and we have found that it provides similar

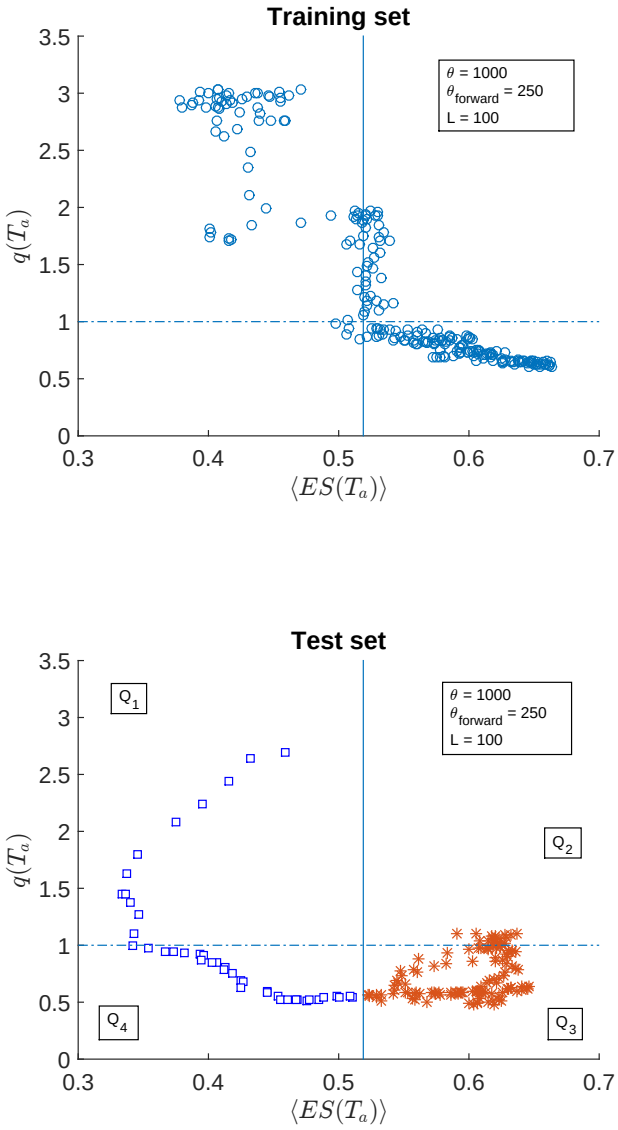

Fig. 6.5 **Partition of data into training (left graphs) and test (right graphs) set (NYSE data set)**. Training sets are used to regress $Y(T_k)$ against $\langle ES \rangle (T_k)$, in order to estimate the coefficents in the logistic regression and therefore identify the regression threshold, shown as a vertical continuous line. The test sets are used to test the forecasting performance of such regression on a subset of data that has not been used for regression; the model predicts $Y(T_k) = 1$ ($q(T_k) > 1$) if $\langle ES \rangle (T_k)$ is greater than the regression threshold, and $Y(T_k) = 0$ ($q(T_k) < 1$) otherwise.

outcomes but worse results in terms of the forecasting performance metrics that we discuss in the next section.

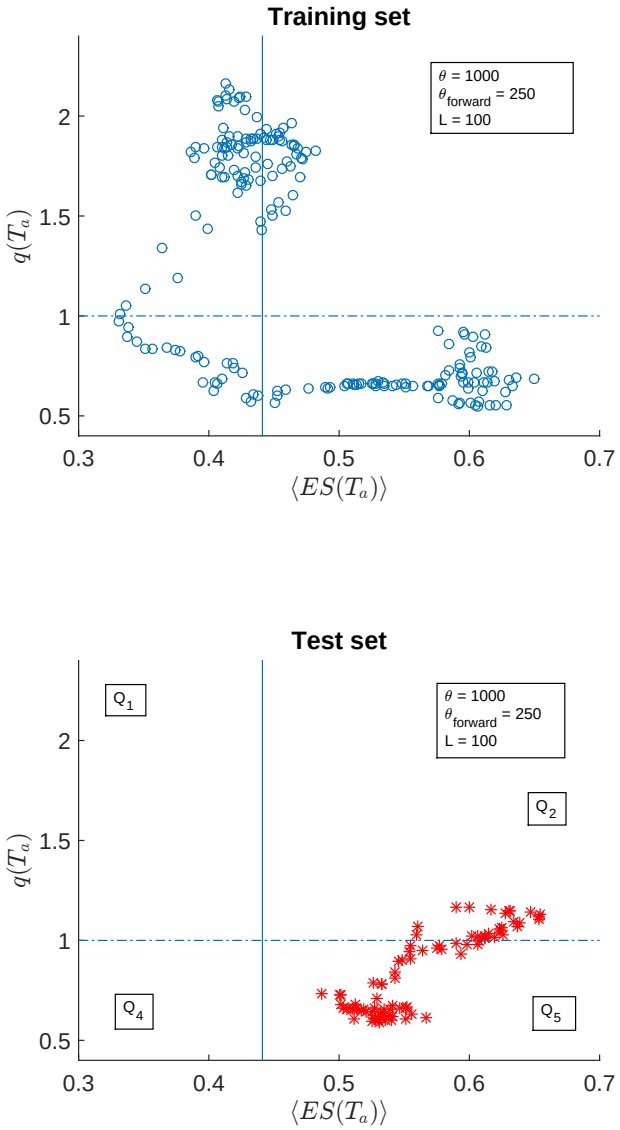

Fig. 6.6 **Partition of data into training (left graphs) and test (right graphs) set (LSE data set)**. Training sets are used to regress $Y(T_k)$ against $\langle ES \rangle(T_k)$, in order to estimate the coefficents in the logistic regression and therefore identify the regression threshold, shown as a vertical continuous line. The test sets are used to test the forecasting performance of such regression on a subset of data that has not been used for regression; the model predicts $Y(T_k) = 1$ ($q(T_k) > 1$) if $\langle ES \rangle(T_k)$ is greater than the regression threshold, and $Y(T_k) = 0$ ($q(T_k) < 1$) otherwise.

6.5.1 Measure of forecasting performance

We here evaluate the goodness of this regression at estimating $Y(T_{n+1})$ given a new observation $\langle ES \rangle(T_{n+1})$. Let us denote with $\{\langle ES \rangle(T_a), q(T_a)\}_{a=1,\ldots,n}$ the set of obser-

vations over which we evaluate our method. With reference to Figs. 6.5-6.6, let us define the number of observed points in each quadrant Q_i ($i = 1,2,3,4$) as $|Q_i|$. In the terminology of classification techniques [15], $|Q_1|$ is the number of True Positive (observations for which the model correctly predicted $Y(T_k) = 1$), $|Q_3|$ is the number of True Negative (observations for which the model correctly predicted $Y(T_k) = 0$), $|Q_2|$ the number of False Negative (observations for which the model incorrectly predicted $Y(T_k) = 0$) and $|Q_4|$ the number of False Positive (observations for which the model incorrectly predicted $Y(T_k) = 1$). We then compute the following measures of quality of classification, that are the standard metrics for assessing the performances of a classification method [15]:

- **Probability of successful forecasting (P^+)** [15]: represents the algorithm probability of a correct prediction, expressed as fraction of observed $\langle ES \rangle(T_k)$ values through which the method has successfully identified the correspondent value of $Y(T_k)$; it is computed as follows:

$$P^+ = \frac{|Q_1| + |Q_3|}{|Q_1| + |Q_2| + |Q_3| + |Q_4|}. \tag{6.9}$$

- **True Positive Rate (TPR)** [15]: it is the probability of predicting $Y(T_k) = 1$, conditional to the fact that the real $Y(T_k)$ is indeed 1 (that is, to predict an increase in volatility when the volatility will indeed increase); it represents the method sensitivity to increase in volatility. In formula:

$$TPR = \frac{|Q_1|}{|Q_1| + |Q_2|}. \tag{6.10}$$

- **False Positive Rate (FPR)** [15]: it is the probability of predicting $Y(T_a) = 1$, conditional to the fact that the real $Y(T_a)$ is instead 0 (that is, to predict an increase in volatility when the volatility will actually decrease). It is also called "1-specificity" [14]. In formula:

$$FPR = \frac{|Q_4|}{|Q_3| + |Q_4|}. \tag{6.11}$$

Overall these metrics provide a complete summary of the model goodness at predicting changes in the market volatility [14].

In order to avoid overfitting we have estimated the metrics above by means of an out-of-sample procedure [14, 15]. We have divided the data set into two time periods, a training set and a test set. In the training set we run the logistic regression and compute the regression threshold r; in the test set we use this r to measure the goodness of the model predictions by computing P^+, TPR and FPR. In Figs. 6.5-6.6 this division is shown for a particular choice of θ and L, for both NYSE and LSE data sets. In this example the percentage of data included in the test set (let us call it f_{test}) is 30%.

Probabilities of successful forecasting P^+ are reported in Tabs. 6.5 and 6.6, for $f_{test} = 30\%$. As we can see P^+ is higher than 50% for all combinations of parameters in NYSE dataset, and in almost all combinations for LSE dataset. Stars mark those values of P^+ that are significantly higher than the same probability obtained by using the most recent value of q instead of $\langle ES \rangle (T_k)$ as a predictor for $q(T_k)$ (let us call P_q^+ such probability). Specifically, we define a null model where variations from such probability P_q^+ are due to random fluctuations only; given n observations, such fluctuations follow a Binomial distribution $B(P_q^+, n)$, with mean nP_q^+ and variance $nP_q^+(1-P_q^+)$. p-values are then calculated by using this null distribution for each combination of parameters. This null hypothesis accounts for the predictability of $q(T_k)$ that is due to the autocorrelation of $q(T_k)$ only; therefore P^+ significantly higher than the value expected under this hypothesis implies a forecasting power of $\langle ES \rangle (T_k)$ that is not explained by the autocorrelation of $q(T_k)$. From the table we can see that P^+ is significant in 12 out of 16 combinations of parameters for NYSE dataset, and in 13 out of 16 for LSE dataset. This means that correlation persistence is a valuable predictor for future volatility, able to outperform forecasting method based on past volatility

trends. These results are robust against changes of f_{test}, as long as the training set is large enough to allow an accurate calibration of the logistic regression. We found this condition satisfied for $f_{test} < 40\%$.

However P^+ does not give any information on the method ability to distinguish between true and false positives. To investigate this aspect we need TPR and FPR. A traditional way of representing both measures from a binary classifier is the so-called "Receiver operating characteristic" (ROC) [200]. In a ROC plot, TPR is plotted against FPR as the discriminant threshold is varied. The discriminant threshold p_{max} is the value of the probability in Eq. 6.8 over which we classify $Y(T_a) = 1$: the higher p_{max} is, the less likely the method is to classify $Y(T_a) = 1$ (in the analysis on P^+ we chose $p_{max} = 0.5$). Ideally, a perfect classifier would yield $TPR = 1$ for all $p_{max} > 0$, whereas a random classifier is expected to lie on the line $TPR = FPR$. Therefore a ROC curve which lies above the line $TPR = FPR$ indicates a classifier that is better than chance at distinguishing true from false positives [14].

As one can see from Figs. 6.7 - 6.8, the ROC curve's position depends on the choice of parameters θ and L. In this respect our classifier performs better for low values of L and θ. This can be quantified by measuring the area under the ROC curve; such measure, often denoted by AUC [14], is shown in Tabs. 6.7-6.8. For both datasets the optimal choice of parameters is $\theta = 500$ and $L = 10$.

Table 6.5 **NYSE dataset: Probability of successful forecasting** P^+, for different combinations of parameters θ and L. Out-of-sample analysis.

		L			
		10	**25**	**50**	**100**
θ	**250**	0.546	0.560*	0.599**	0.539**
	500	0.704**	0.695**	0.658**	0.605**
	750	0.634*	0.585	0.539	0.708*
	1000	0.704*	0.7638**	0.839**	0.860

$^{**}p < 0.001$, $^{*}p < 0.01$,

Table 6.6 **LSE dataset: Probability of successful forecasting** P^+, for different combinations of parameters θ and L. Out-of-sample analysis.

		L			
		10	**25**	**50**	**100**
θ	**250**	0.616**	0.645**	0.612**	0.568**
	500	0.652**	0.635**	0.598**	0.393
	750	0.651**	0.560**	0.453**	0.412
	1000	0.544**	0.573**	0.706**	0.689

$^{**}p < 0.001$, $^{*}p < 0.01$,

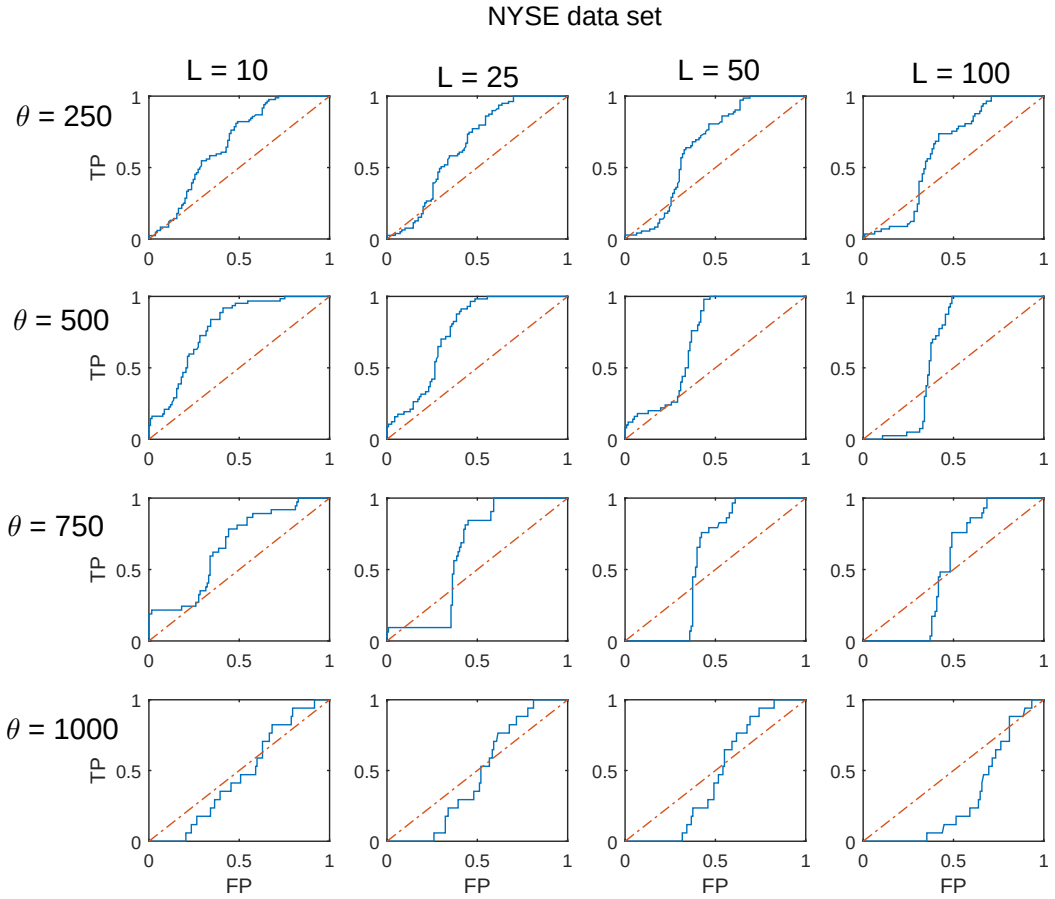

Fig. 6.7 **Receiver operating characteristic (ROC) curve for the NYSE dataset**. True positive rate (TPR) against False positive rate (FPR) as the discriminant threshold p_{max} of the classifier is varied, for each combination of parameters θ and L in the NYSE dataset. The closer the curve is to the upper left corner of each graph, the better is the classifier compared to chance.

Table 6.7 **NYSE dataset: Area under the curve (AUC)**, measured from the ROC curve in Fig. 6.7. Values greater than 0.5 indicate that the classifier performs better than chance.

		L			
		10	**25**	**50**	**100**
θ	**250**	0.669	0.652	0.655	0.616
	500	0.775	0.753	0.710	0.625
	750	0.663	0.6220	0.574	0.520
	1000	0.467	0.470	0.462	0.314

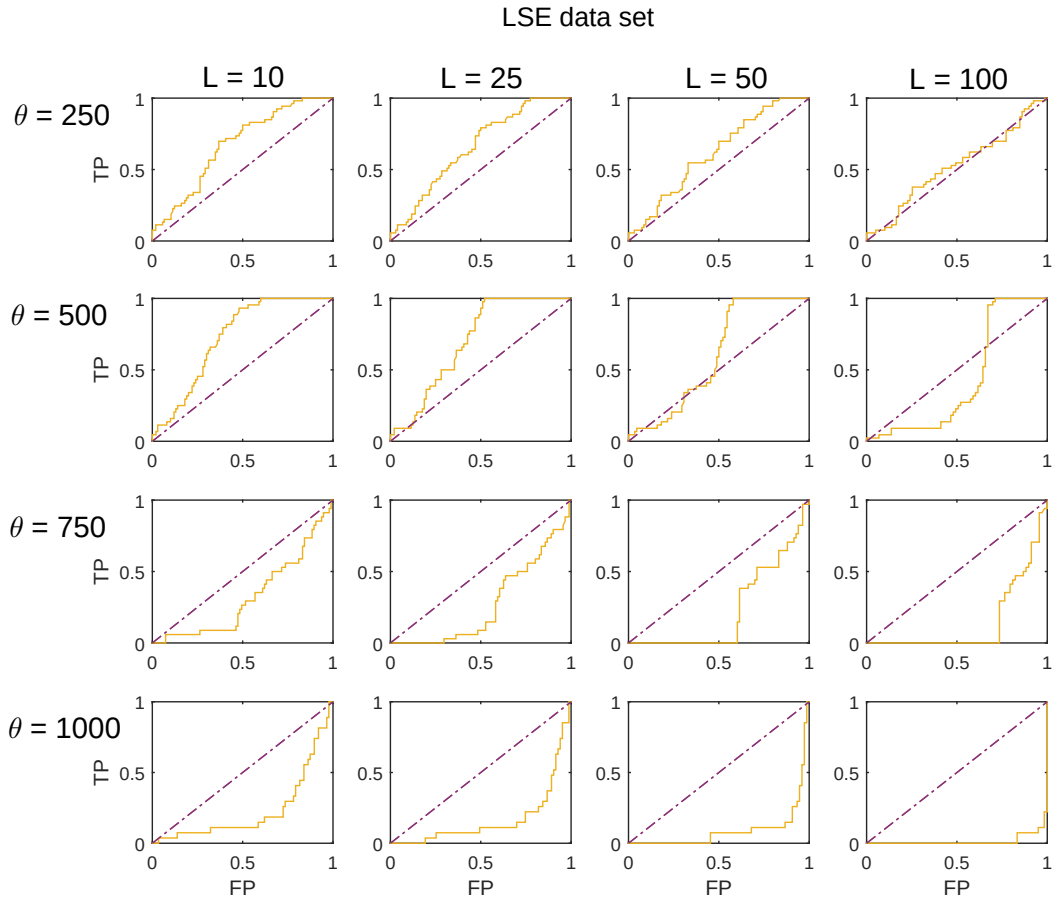

Fig. 6.8 **Receiver operating characteristic (ROC) curve for the LSE dataset**. True positive rate (TPR) against False positive rate (FPR) as the discriminant threshold p_{max} of the classifier is varied, for each combination of parameters θ and L in the LSE dataset. The closer the curve is to the upper left corner of each graph, the better is the classifier compared to chance

Table 6.8 **LSE dataset: Area under the curve (AUC)**, measured from the ROC curve in Fig. 6.7. Values greater than 0.5 indicate that the classifier performs better than chance.

		L			
		10	**25**	**50**	**100**
θ	**250**	0.673	0.658	0.618	0.524
	500	0.727	0.700	0.602	0.431
	750	0.324	0.274	0.234	0.148
	1000	0.233	0.168	0.0918	0.0160

6.5.2 Temporal evolution of forecasting performance

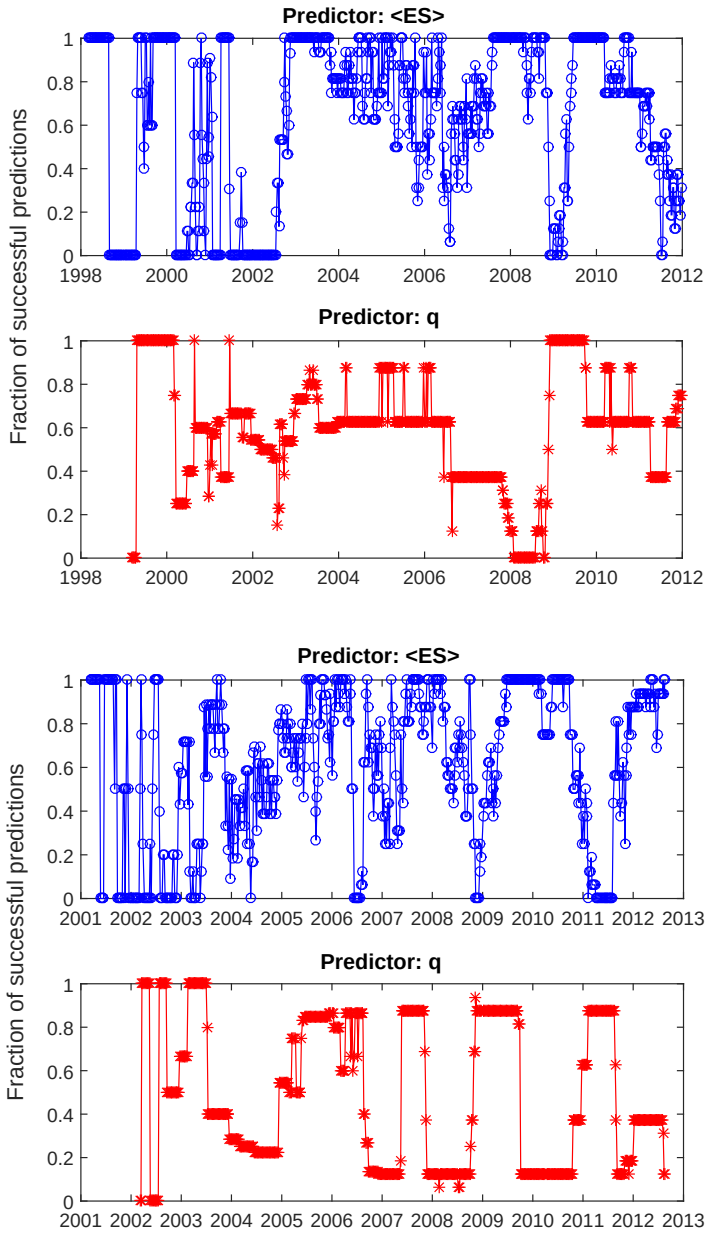

Fig. 6.9 **Fraction of successful predictions as a function of time**. NYSE (right graphs) and LSE dataset (left graphs). Forecasting is based on logistic regression with predictor $\langle ES(T_k)\rangle$ (top graphs) and most recent value of $q(T_k)$ (bottom graphs). Horizontal lines represent the average over the entire period.

In this section we discuss how the forecasting performance changes at different time periods. In order to explore this aspect we have counted at each time window T_k the number $N^+(T_k)$ of $Y(T_k)$ predictions (out of the 16 predictions corresponding to

as many combinations of θ and L) that have turned out to be correct; we have then calculated the fraction of successful predictions $n^+(T_k)$ as $n^+(T_k) = N^+(T_k)/16$. In this way $n^+(T_k)$ is a proxy for the goodness of our method at each time window. For each combination of parameters the model is calibrated by using the entire time period as training set, therefore this amounts to an in-sample analysis.

In Fig. 6.9 we show the fraction of successful predictions for both NYSE and LSE data sets (blue dots). For comparison we also show the same measure obtained by using the most recent value of $q(T_k)$ as predictor (red stars); as in Section 6.5, it represents a null model that makes prediction by using only the past evolution of $q(T_k)$. As we can see, both predictions based on $\langle ES \rangle(T_k)$ and on past values of $q(T_k)$ display performances changing in time. In particular $n^+(T_k)$ drops just ahead of the main financial crises (the market downturn in March 2002, 2007-2008 financial crisis, Euro zone crisis in 2011); this is probably due to the abrupt increase in volatility that occurred during these events and that the models took time to detect. After these drops though performances based on $\langle ES \rangle(T_k)$ recover much more rapidly than those based on past value of $q(T_k)$. For instance in the first months of 2007 our method shows quite high $n^+(T_k)$ (more than 60% of successful predictions), being able to predict the sharp increase in volatility to come in 2008 while predictions based on $q(T_k)$ fail systematically until 2009. Overall, predictions based on dependence structure persistence appear to be more reliable (as shown by the average $n^+(T_k)$ over all time windows, the horizontal lines in the plot) and faster at detecting changes in market volatility.

6.6 Summary

In this chapter we have demonstrated that there is a deep interplay between market volatility and the rate of change of the dependence structure. In particular the latter can be used to forecast valuable information about future values of the former, providing a

useful tool for risk and portfolio management. This interplay is better highlighted when filtering based on Planar Maximally Filtered Graphs is used to estimate the dependence structure persistence. We have proved the forecasting power of this tool by means of out-of-sample analyses on two different stock markets, showing that it can outperform predictions based on past market volatility trends. Moreover we have measured True and False positive rates to identify an optimal region of the parameters in terms of forecasting reliability. The advantage of our approach over traditional econometrics tools, such as multivariate GARCH and stochastic covariance models, is the "top-down" methodology that treats correlation matrices as the fundamental objects, allowing to deal with many assets simultaneously; in this way the curse of dimensionality, that prevents e.g. multivariate GARCH to deal with more than few assets, is avoided.

Aside from the applications, our results shed new light into the dynamic of correlation evolution. Topology of Planar Maximally Filtered Graphs depends on the ranking of the $N(N-1)/2$ pairs of cross-correlations; therefore an increase in the rate of change in PMFGs topology points out a faster change of this ranking. Our result indicates that such increase is typically followed by a rise in the volatility, whereas decrease are followed by drops. A possible interpretation of this is related to the dynamics of risk factors in the market. Indeed higher volatility in the market is associated to the emergence of a (possibly new) risk factor that makes the whole system more vulnerable; such transition could be anticipated by a quicker change of the correlation ranking, triggered by the still emerging factor and revealed by the dependence structure persistence. Such persistence can therefore be a powerful tool for monitoring the emergence of new risks, valuable for a wide range of applications, from portfolio management to systemic risk regulation. Moreover this interpretation would open interesting connections with those approaches to systemic risk that makes use of Principal Component Analysis, monitoring the emergence of new risk factors by means of spectral methods [197, 198].

From these analyses we find evidence that non-stationarity of correlation is a fundamental aspect, which has deep interplay with the evolution of risk. In the next

chapter we investigate the issue of correlation dynamics from a slightly different perspective, namely its degree of non-linearity, and we analyse how non-stationarity affects its importance.

Chapter 7

Multiplex on correlation-based networks

In this chapter we measure the degree of non-linearity in the dependence structure through network filtering. To this, end we apply for the first time the multiplex framework to correlation-based networks. Such a framework allows us to quantify the degree of dissimilarity among PMFGs constructed from different measures of dependence, both linear and non-linear. We find evidence of deep differences among these measures, which indicates a degree of non-linearity in the dependence structure. By using a rolling time window analysis, we also show how this non-linearity is time dependent. The implications of these findings for risk management are discussed as well. The results and analyses presented in this chapter are based on a paper that has been submitted to a peer-reviewed scientific journal in 2016.

7.1 Introduction

So far we have used only Pearson coefficient as a measure of dependence. The reason for this choice is due to the great popularity of this measure: in the vast majority of applications it is indeed the only measure of dependence used [110]. However, as we have discussed in Chapter 2, Pearson is an optimal measure only if the two variables

of interest are drawn from a multivariate normal distribution. If this is not the case, non-linearity arises in the relation between the two random variables, and Pearson is no longer a natural choice.

Some works [49, 50] have indeed demonstrated that non-linearity is a distinctive feature of financial correlation. There are manifestations of non-linearity which have been misinterpreted as signatures of non-stationarity [47]. The main limitation of these studies is however their focus on pairwise relations. To the best of our knowledge, the literature has overlooked the issue of assessing the degree of non-linearity in terms of the entire dependence structure and its implication for risk estimation at the market level.

We here propose an approach based on correlation-based networks to investigate this issue. There are measures of dependence alternative to Pearson, such as Kendall [107] and Tail dependence [205], which are able to capture part of non-linearity [110]. Since correlation-based networks can be constructed from any similarity measure, our strategy consists in comparing the topology of Pearson-based network with those of networks built from non-linear dependence measures. The degree of difference would quantify the extent of non-linearity in the dependence structure of financial returns. Some works have been devoted to study correlation-based networks with measures of dependence different from Pearson coefficient [71, 72, 206, 207]; however, these different networks have never been compared systematically so far.

In order to perform systematically such network comparison we make use of multiplex networks [84, 85, 208], namely a set of tools designed to quantify the interplay among two or more networks (called layers) defined on the same set of nodes. They have been growing in popularity in the last decade across a variety of disciplines, as they allow more refined and complete network analyses than traditional, single-layer network theory tools [84].

The original contributions of this chapter are the following:

- We apply for the first time the multiplex analysis to correlation-based networks. In particular we use the degree of dissimilarity between networks computed from different dependence measures (Kendall [107], Tail [205] and Partial correlation [183]) as a proxy for the degree of non-linearity in the dependence structure. We find that the influence of non-linearity has changed over the last 22 years, increasing in particular during turbulent market periods.

- We analyse how different dependence measures assess nodes centrality in the dependence structure, by using appropriate multiplex metrics. We show that the result is strongly dependent on the dependence measure; moreover, these differences are time dependent. These results indicate that Pearson coefficient alone is not sufficient to monitor the evolution of financial dependencies.

This chapter is organized as follows. In Section 7.2 we introduce the concept of multiplex, we provide a brief summary of the literature and describe the multiplex metrics we will use in the chapter. In Section 7.3 we introduce the non-linear dependence measures which we have used as layers in the multiplex analysis. In Section 7.4 we describe the data set which we have used in our analyses. In Section 7.5 we show the results of the application of multiplex to the data set, as well as discuss their implications.

7.2 Multiplex: a brief introduction

The idea of analysing multiple layers of interaction was introduced initially in the context of social networks, within the theory of frame analysis [209]. The importance of considering multiple types of human interactions has been more recently demonstrated in different social networks, from terrorist organizations [85] to online communities; in all these cases, multilayer analyses unveil a rich topological structure [87], outperforming single-layer analyses in terms of network modeling and prediction as well [88, 89, 210]. In particular multilayer community detection in social networks has

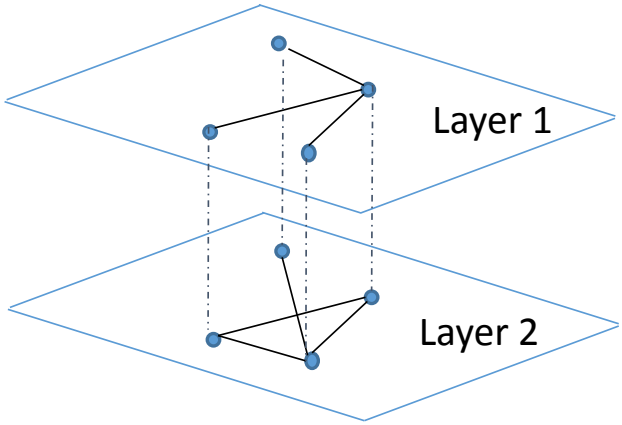

Fig. 7.1 **Schematic representation of a two-layers multiplex**.

been shown to be more effective than single-layer approaches [211]; similar results have been reported for community detection on the World Wide Web [90, 91] and citation networks [92]. In the context of electrical power grids, multilayer analysis has provided important insight into the role of synchronisation in triggering cascading failures [212, 213]. Similarly, the analyses on transport networks have highlighted the importance of a multilayer approach to optimise the system against nodes failures, such as flights cancellation [214].

In the context of Economic networks, multiplex analyses have been applied to analyse the World Trade Web [215], where the commodity-specific trade network is compared to the aggregate trade network. Moreover, they have been extensively used in the context of systemic risk, where graphs are used to model interbank and credit networks [216, 93–95]. In [93] it has been shown that focusing on a single layer underestimates the total systemic risk of the Mexican banking system by up to 90%. Moreover, a non-linear effect makes the total systemic risk higher than the sum of systemic risks across layers [93]. A similar behaviour is observed in an agent-based model presented in [94], where it is also introduced a novel "systemic importance" measure based on multi-layer networks. In [95] this non-linear effect is analysed as a function of the interaction strength among layers, revealing the existence of a critical

strength above which the use of multiplex is crucial for risk estimation. Despite this interest in Economics, to the best of our knowledge multiplexes have never been applied to correlation-based networks. With this chapter we aim to bridge this gap.

Let us here introduce formally some notation on multiplex [85]. A multiplex is a set of networks defined on a common set of nodes [84], as shown in Fig. 7.1 for a simple case of four nodes and two layers. Although most of the metrics we will use can be easily generalised to the weighted case, we will focus on unweighted networks.

Formally, let us define a M-dimensional array of adjacency matrices:

$$\mathscr{A} = \{a^{[1]}, a^{[2]}, \ldots, a^{[M]}\} \ , \tag{7.1}$$

where $a_{ij}^{[\alpha]}$ indicates the presence ($a_{ij}^{[\alpha]} = 1$) or absence ($a_{ij}^{[\alpha]} = 0$) of links between nodes i and j on α-th network. This array is called multiplex and is denoted by symbol $\mathscr{M}$. Furthermore, we denote by $K^{[\alpha]} = \frac{1}{2}\sum_{ij} a_{ij}^{[\alpha]}$ the number of edges on layer α, and by $K = \frac{1}{2}\sum_{i,j}\left[1 - \prod_\alpha (1 - a_{ij}^{[\alpha]})\right]$ the number of pairs of nodes which are connected by one edge on at least one layer. Notice that since the network at each layer is a PMFG, then we necessarily have $K^{[\alpha]} = 3(N-2)$ for every α as discussed in Chapter 3.

To have a first, global measure of similarity among the different layers the mean edge overlap has been proposed [85], which is defined as the average number of layers on which an edge between two randomly chosen nodes i and j exists:

$$\langle O \rangle = \frac{1}{2K}\sum_{i,j}\sum_{\alpha} a_{ij}^{[\alpha]} \ . \tag{7.2}$$

The mean edge overlap is a measure of how similar the multiplex layers are; indeed, $\langle O \rangle = 1$ only when all the M layers are identical, i.e. $A^{[\alpha]} \equiv A^{[\beta]} \ \ \forall \ \ \alpha, \beta = 1, \ldots, M$, while $\langle O \rangle = 0$ if no edge is present in more than one layer at the same time.

A quantity related to the overlapping degree is the fraction of edges of layer α which do not exist on any other layer, which we can compute as follows:

$$U^{[\alpha]} = \frac{1}{2K^{[\alpha]}} \sum_{i,j} a_{ij}^{[\alpha]} \prod_{\beta \neq \alpha} \left(1 - a_{ij}^{[\beta]}\right) \quad . \tag{7.3}$$

$U^{[\alpha]}$ is close to zero only when almost all the edges of layer α are also present on at least one of the other $M-1$ layers. In this sense, it quantifies the peculiarity of layer α with respect to the other layers in the multiplex.

$\langle O \rangle$ and $U^{[\alpha]}$ quantify the amount of links shared among the layers. If we want to obtain information regarding the degree variability across layers (i.e. to what extent the layers agree on the importance of nodes, as measured by the degree defined in Chapter 3), different measures are needed. When it comes to only two layers, a simple approach is to measure the Pearson coefficient between the degree distributions of the two layers [85]:

$$\rho_{\alpha_1 \alpha_2}^{deg} \equiv \rho(\{k_i^{\alpha_1}\}, \{k_i^{\alpha_2}\}) \quad , \tag{7.4}$$

where $\{k_i^{\alpha_1}\}$ and $\{k_i^{\alpha_2}\}$ are the degree sequences respectively on layer α_1 and α_2 [85]. The limit of this approach is that it is intrinsically pairwise; moreover it does not provide information at the node level. To overcome this limitation, two measures have been proposed [85], namely the overlapping degree:

$$o_i = \sum_{\alpha} k_i^{[\alpha]} \quad , \tag{7.5}$$

and the multiplex participation coefficient:

$$P_i = \frac{M}{M-1} \left[1 - \sum_{\alpha} \left(\frac{k_i^{[\alpha]}}{o_i}\right)\right] \quad . \tag{7.6}$$

The overlapping degree is just the total number of edges incident on node i at any layer, whereas the multiplex participation coefficient quantifies the dispersion across the layers

of the edges incident on node i. In fact, $P_i = 0$ if the edges of i are concentrated on exactly one of the M layers, while $P_i = 1$ if the edges of i are uniformly distributed across the M layers, i.e. when $k_i^{[\alpha]} = \frac{o_i}{M} \ \ \forall \ \ \alpha$ (in which case i is a truly multiplex node, as its topological role is somehow consistent across the layers). The scatter plot of o_i and P_i for a given multiplex is called *multiplex cartography* and has been used as a synthetic graphical representation of the overall heterogeneity of node roles observed in a multiplex.

We can explore deeper single nodes' role across the layers by means of the concept of multilink and multidegree. Let us define the vector $\vec{m} = (m_1, m_2, ..., m_M)$, where each m_α can take only two values $\{1, 0\}$. We say that a pair of nodes i, j has a multilink $\vec{m}$ if they are connected only on those layers α for which $m_\alpha = 1$ in $\vec{m}$ [84]. The information on the M adjacency matrices a_{ij}^α can then be aggregated in the multiadjacency matrix $A_{ij}^{\vec{m}}$: $A_{ij}^{\vec{m}} = 1$ if and only if the pair i, j is connected by a multilink $\vec{m}$. Formally [84]:

$$A_{ij}^{\vec{m}} \equiv \prod_{\alpha=1}^{M} [a_{ij}^\alpha m_\alpha + (1 - a_{ij}^\alpha)(1 - m_\alpha)] \ . \tag{7.7}$$

From the multiadjacency matrix we can define the multidegree $\vec{m}$ of a node i, as the number of multilinks $\vec{m}$ connecting i:

$$k_i^{\vec{m}} = \sum_j A_{ij}^{\vec{m}} \ . \tag{7.8}$$

This measure allows us to calculate e.g. how many edges node i has on layers 1 and 3 only ($k_i^{\vec{m}}$ choosing $m_1 = 1$, $m_\alpha = 0 \ \ \forall \ \ \alpha \neq 1, 3$), integrating the global information provided by $U^{[\alpha]}$.

7.3 Beyond Pearson coefficient: non-linear measures of dependence

We here introduce and discuss different measures of dependence which we will analyse within the multiplex set-up, along with the Pearson coefficient.

7.3.1 Spearman and Kendall correlation

To overcome the flaws of Pearson coefficient with non-linear dependences two alternative measures can be used, Spearman [108] and Kendall correlations [107]. They are called ordinal correlations, as they look at the ranking of the observations rather than the values: this feature allows them to capture any monotonic dependence, unlike Pearson that can detect only linear relations. Another advantage is that Spearman and Kendall correlations are both non-parametric, that is they can be used without making any assumption on the variables distribution.

Spearman correlation $\rho_S(i,j)$ is defined as the Pearson coefficient calculated on the ranking of the observations, rather than on the observations [108]. Let us assume we have a set of T observations for log-returns of two stocks, $r_i(t)$ and $r_j(t)$, with $t = 1, ..., T$. Defining $R^{r_i(t)}$ and $R^{r_j(t)}$ as the rankings of observations $r_i(t)$ and $r_j(t)$, and $d_t = R^{r_i(t)} - R^{r_j(t)}$, Spearman correlation can be written as [110]:

$$\rho_S(i,j) = 1 - \frac{6\sum_t d_t^2}{T(T^2-1)} \ , \tag{7.9}$$

where T is the total number of observations.

Kendall correlation is defined as follows [107]:

$$\tau(i,j) = \frac{\sum_{t=1}\sum_{s=t+1} d_{ts}^i d_{ts}^j}{\sqrt{[\frac{1}{2}T(T-1) - n^i][\frac{1}{2}T(T-1) - n^j]}} \ , \tag{7.10}$$

where now $d_{ts}^i \equiv sgn(R^{r_i(t)} - R^{r_i(s)})$, and n^i is the number of tied pairs (i.e. cases where $d_{ts}^i = 0$). The numerator counts the number of concordant pairs (i.e. pairs of

observations such that d^i_{ts} and d^j_{ts} have equal signs) minus the number of discordant pairs; the denominator is a normalization factor that takes into account possible tied pairs [103]. A weighted version of Kendall correlation has been proposed [217], similarly to the Pearson case in Eq. 2.19:

$$\tau^w(i,j) = \sum_{t=1} \sum_{s=t+1} w_{t,s} d^i_{ts} d^j_{ts} \ , \tag{7.11}$$

with

$$w_{t,s} = w_0 \exp\left(\frac{t-T}{\theta}\right) \exp\left(\frac{s-T}{\theta}\right) \ . \tag{7.12}$$

If $r_i(t)$ and $r_j(t)$ are drawn from a multivariate normal distribution, it can be shown that there is a direct relation between Kendall and Pearson coefficient [218]:

$$\tau(i,j) = \frac{2}{\pi} \arcsin(\rho_{ij}) \ . \tag{7.13}$$

Hence in the multivariate normal case $\tau(i,j)$ and ρ_{ij} carry the same information on the dependence between the two random variables. In particular, since the relation in Eq. 7.13 is an increasing function, the topology structure obtained from correlation-based networks such as MST or PMFG would not change by using Kendall instead of Pearson, since such structure depends ultimately on the ranking of pairwise dependences [64]. This result holds not only for multivariate normal distributions, but it is true for the broader class of elliptical distributions [218]. Therefore, any difference between the topology of Pearson-based and Kendall-based correlation-based networks would be an indication of non-linearity in the dependence structure.

7.3.2 Tail-dependence

Despite their advantages, Kendall and Spearman coefficients tend to underestimate risk because are less sensitive to outliers than Pearson [110]. It is possible to define a

dependence measure that describes co-movements only in the tails of the multivariate distribution. This tail dependence is of great interest for applications, as large deviations from the mean are the main focus in Risk Management: if two assets tend to correlate in the tails differently from how they behave closer to the mean, then this difference must be quantified and be taken into account [112].

Following the notation of [205], let us call G and H the marginal cumulative distributions of returns r_i and r_j. The upper-tail dependence between r_i and r_j is defined as the probability of one variable taking a value very far in the upper tail region, conditional to the other one taking a value very far in the upper tail region as well; formally, this is expressed as follows:

$$\lambda_U(i,j) = \lim_{u \to 1^-} P(r_i > G^{-1}(u) | r_j > H^{-1}(u)) \ . \tag{7.14}$$

Similarly, the lower-tail dependence is defined as:

$$\lambda_L(i,j) = \lim_{u \to 0^+} P(r_i < G^{-1}(u) | r_j < H^{-1}(u)) \ . \tag{7.15}$$

It can be shown that multivariate normal distributions are upper and lower tail-independent, that is $\lambda_U(i,j) = \lambda_L(i,j) = 0$ [49]. In this sense, measures of tail-dependence are again measures of non-linearity in the empirical multivariate distribution. Other examples of tail-independent distributions are multivariate hyperbolic and logistic distributions, whereas multivariate t-student and α-stable distributions are upper and lower tail-dependent. It has been shown that empirical equity returns are compatible with upper tail-independence, but display a strong lower tail-dependence [49]. This means that when control of risk is most needed (during periods of negative returns) the dependence structure is maximally distant from the picture provided by Pearson coefficient [113].

It is possible to estimate tail-dependence by means of non-parametric estimators as follows [219]:

$$\hat{\lambda}_U(i,j,p_1,p_2) = \frac{\sum_{t=1}^{\theta} \mathbb{1}_{\{G(r_i(t))>p_1 \wedge H(r_j(t))>p_2\}}}{\sum_{t=1}^{\theta} \mathbb{1}_{\{G(r_i(t))>p_1 \vee H(r_j(t))>p_2\}}} \ , \tag{7.16}$$

$$\hat{\lambda}_L(i,j,p_1,p_2) = \frac{\sum_{t=1}^{\theta} \mathbb{1}_{\{G(r_i(t))<p_1 \wedge H(r_j(t))<p_2\}}}{\sum_{t=1}^{\theta} \mathbb{1}_{\{G(r_i(t))<p_1 \vee H(r_j(t))<p_2\}}} \ , \tag{7.17}$$

where p_1 and p_2 are two parameters representing the percentiles above (below) which an observation is considered upper (lower) tail. These estimators converge to the corresponding tail-dependences in Eq. 7.14-7.15 in the limit $p_{1,2} \to 1^-$ for $\hat{\lambda}_U$ and $p_{1,2} \to 0^+$ for $\hat{\lambda}_L$. In practice the optimal choice of $p_{1,2}$ should be a trade-off between such convergence and the availability of data in the tails.

7.3.3 Partial correlation

Significant dependence between two random variables does not imply a causal relation between them; rather, it can be due to the presence of a third, common factor that influences both variables. Partial correlation [183] can be used to distinguish between direct and indirect relationships. It is a measure of dependence that quantifies to what extent r_i and r_j are dependent, after taking into account (and subtracting) the influence of a third variable r_k. Specifically, the Partial correlation between assets r_i and r_j based on r_k, $\rho_{ij|k}$, is the Pearson correlation between the residuals of r_i and r_j obtained after linear regression against r_k [183]. It can be written in terms of Pearson correlation as follows [71]:

$$\rho_{ij|k} = \frac{\rho_{ij} - \rho_{ik}\rho_{jk}}{\sqrt{[1-\rho_{ik}^2][1-\rho_{jk}^2]}} \ . \tag{7.18}$$

This measure represents the amount of correlation between r_i and r_j that is left once the influence of r_k is subtracted.

If the aim is to quantify the influence of r_k on r_i and r_j, $\rho_{i,j|k}$ is actually not a good measure: a low value of $\rho_{i,j|k}$ could be due to either a strong influence of r_k or to a weak

relation between r_i, r_j, regardless of r_k. For this reason in [71] the authors suggest to describe the relation among r_i, r_j and r_k by defining the correlation influence of r_k on the pair r_i, r_j as:

$$d(i,j|k) = \rho_{ij} - \rho_{ij|k} \tag{7.19}$$

In this way $d(i,j|k)$ is large when a significant fraction of correlation between r_i and r_j is due to the influence of r_k. It is possible to translate this measure into a direct measure of influence between r_i and r_k - called "average correlation influence" - by averaging it over a set of variables r_j [71]:

$$d(i|k) = \langle d(i,j|k) \rangle_{j \neq i,k} \tag{7.20}$$

7.4 Data set: non-continuously traded stocks

For the multiplex analysis we have used a data set consisting of daily prices of $N_{tot} = 1004$ US stocks, traded in the period between 03/01/1993 and 26/02/2015. Each stock in the dataset has been included in S&P500 at least once in the period considered, hence they provide a representative picture of the US stock market over 22 years, covering all the 10 industries listed in the Industry Classification Benchmark ICB [192].

It is important to note that most of stocks in this set are not traded over the entire period. In the analyses we have discussed in the previous chapters we had selected only those stocks which were continuously traded over the analysed period: this constraint was necessary because the network analyses required a fixed set of nodes in time. However this condition is no longer necessary in the multiplex analysis, as we only need the set of nodes to be fixed across layers at the same time window. The advantage of this new data set is that the set of stocks is more representative of the market, avoiding any "survival bias".

The dynamic analysis has been performed by using the rolling time window set-up described in Chapter 2. Time windows length has been chosen equal to $\theta = 1000$ trading days (about 4 years), with a shift of $dT = 23$ trading days (one month), adding up to 200 time windows. On each window 4 dependence matrices $N \times N$ have been calculated, corresponding to Pearson, Kendall, Tail and Partial correlation (see Section 7.5). Since the number of active stocks changes with time, dependence matrices at different times have different number of stocks, as shown in Fig. 7.2. In the figure is also shown the ICB industry composition of this data set in each time window, confirming that we have a representative sample of all market throughout the period. We have verified that the results we are discussing in the following are robust against change of θ and dT.

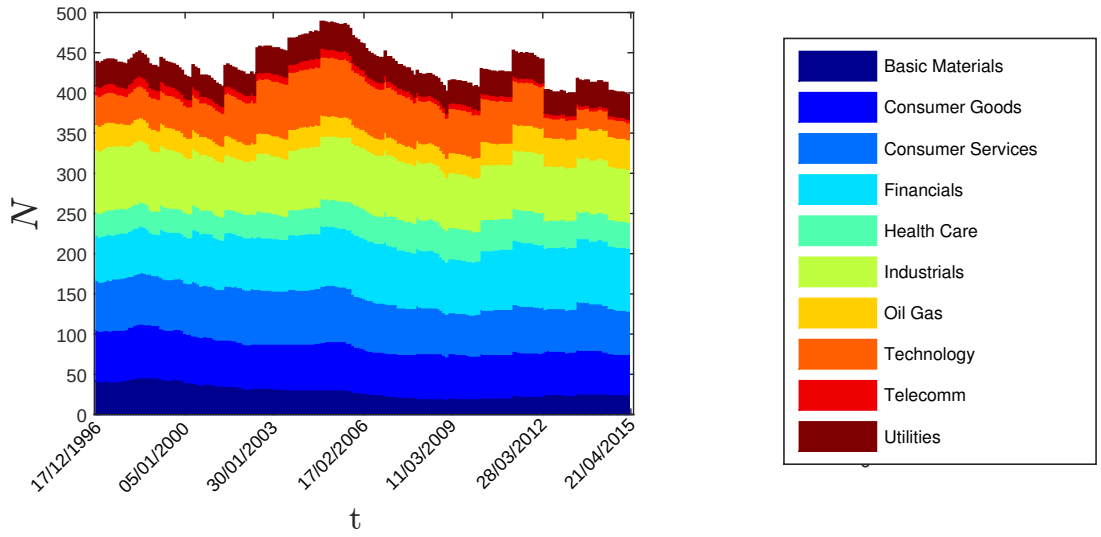

Fig. 7.2 **Number of stocks that are continuously traded in each time window together with their partition in terms of ICB industries.**

7.5 Multiplex on correlation-based networks

In this section we apply the multiplex tool to the data set we have introduced in Section 7.4. For each time window we construct a multiplex composed of 4 layers; each layer is a PMFG built from a different dependence measure. As discussed, we focus on the following measures:

1. **Layer 1: Pearson correlation** ρ_{ij}, computed from the weighted estimator in Eq. 2.19 in Chapter 2;

2. **Layer 2: Kendall's tau** $\tau(i,j)$, computed from the weighted estimator in Eq. 7.11;

3. **Layer 3: Empirical lower tail copula** $\lambda_U(i,j)$, computed from estimator in Eq. 7.17. We focus on the lower tail as it is of interest for risk management applications; moreover, as we discussed, evidences of non-linearity have been found mostly on the lower tails [49]. As for the values of parameters $p_{1,2}$ in Eq. 7.17 defining the lower tail threshold, we have chosen $p_1 = p_2 = 0.1$ (i.e. we consider tail every observation below the 10th percentile), as a trade-off between the need of statistic and the interest in extreme events.

4. **Layer 4: Average correlation influence** $d(i|j)$, computed from Eq. 7.20. It is worth noting that, unlike the other layers, $d(i|j)$ provides a direct relation between assets (as in general $d(i|j) \neq d(j|i)$). This requires an adaptation of the PMFG algorithm that is able to deal with asymmetric relations: we have followed the approach suggested in [71], that rules out double links between nodes. In the rest of the paper we refer to this layer as "Partial layer", even though strictly speaking we are analysing the Correlation influence based on Partial correlation.

We do not consider a layer of Spearman correlation, as we found it provides very similar results to Kendall in terms of the metrics which we will introduce in this section.

7.5.1 A global look at non-linearity: edge overlap

To have a first, global measure of similarity among the different layers we have measured the mean edge overlap defined in Eq. 7.2. At each time window we have calculated $\langle O \rangle$. This calculation has been performed both on the whole 4 layers multiplex and on pairwise multiplexes (i.e. those multiplexes obtained by considering only one pair of layers at a time, totalling $\frac{6 \cdot 5}{2!} = 6$ multiplexes). We will refer to the latters with $\langle O \rangle_{\alpha_1 \alpha_2}$, where $\alpha_{1,2}$ identifies the pair of layers.

In Fig. 7.3 the results are shown. On each graph we have added vertical lines to highlight the main financial crises since 1997. The 4 layers mean edge overlap $\langle O \rangle$ displays a quite dynamic pattern: in particular a steep rise starts at the end of 2000 reaching its peak in 2002, just before the same year market downturn that marks a first, sensitive decrease. Pre-2000 values are not reached until 2005 though, when a steep decline leads in 2007 the mean edge overlap to the lowest value since 1998. Interestingly this decline coincides with the second phase of the housing bubble, and terminates in the middle of 2007, when the credit crunch starts to spread globally. A second, even steeper drop occurs with the Lehman Brothers default. After it the measure appears more stable and weakly increasing, especially at the end of 2014. A part from the aforementioned 2002 and 2007-08 crises, only the Russian crisis in 1998 seems to trigger a sensitive variation.

The pairwise analyses give an insight into the contribution of each pair of layers to the global picture described above. Evolutions of $\langle O \rangle_{\alpha_1 \alpha_2}$ for Pearson-Kendall, Pearson-Tail, Kendall-Tail and Tail-Partial look qualitatevily pretty similar to each other and to the 4 layers graph. This means that the main movements in the latter signal are due to changes in the contribution of both non-linearity (Pearson-Kendall pair) and extreme events (pairs with Tail layer). However, while the pairs including Tail appear smoother, the Pearson-Kendall graph displays a much higher sensitivity to financial crises (even to the 1997 Asian crisis, not evident in the 4 layers case). This implies that non-linearity becomes more and more important during financially turbulent periods. Finally, pairs

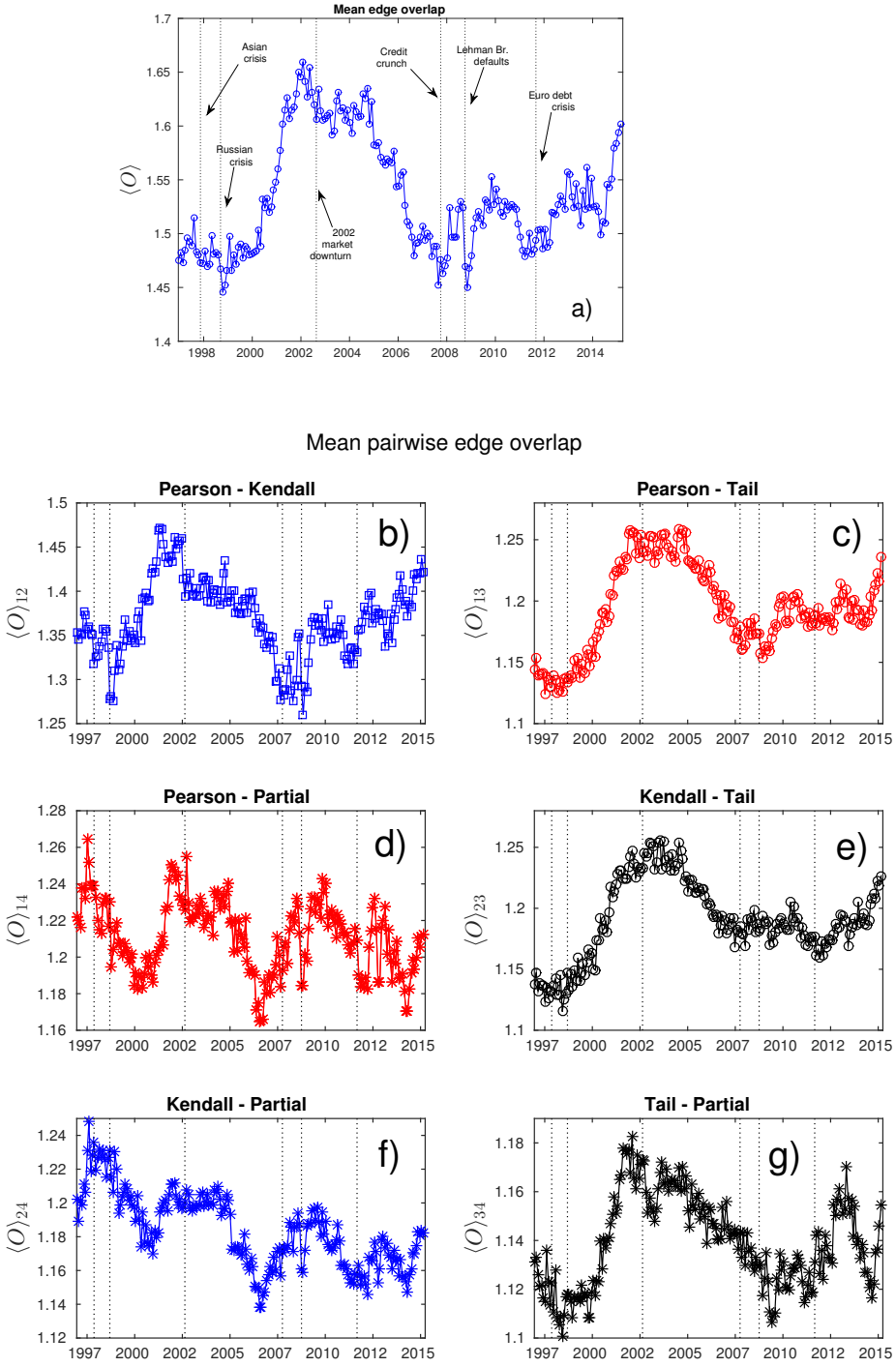

Fig. 7.3 **Mean edge overlap evolution in time**. Upper graph: mean edge overlap $\langle O \rangle$ for the whole (4 layers) multiplex, calculated at each time window. Vertical lines highlight the main financial crises since 1997. Bottom: Mean edge overlap $\langle O \rangle_{\alpha_1 \alpha_2}$ calculated on multiplex made of two layers α_1 and α_2.

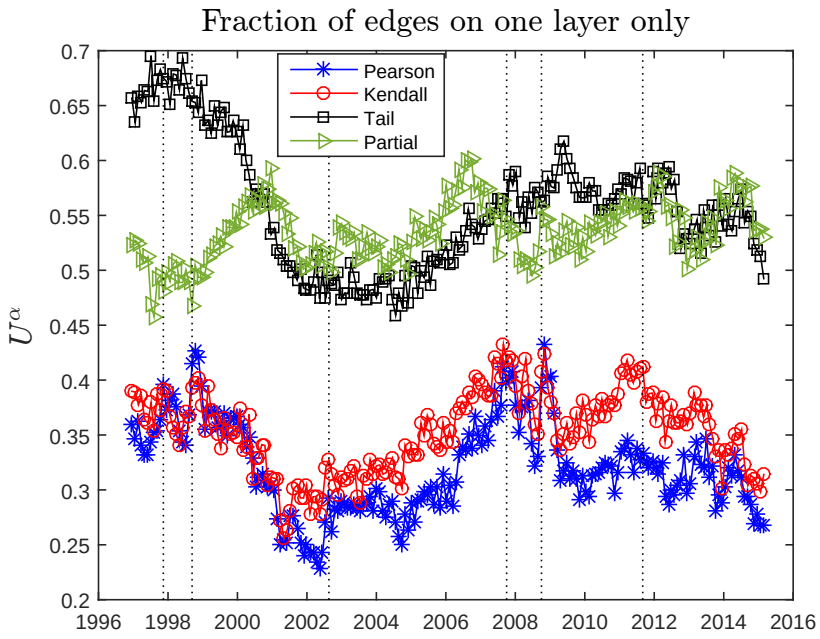

Fig. 7.4 **Fraction of edges that exist only on one layer: evolution in time**. For each layer α, the number of edges that exist only on that layer has been calculated (U^α) and then normalized by the total number of edges in each layer (N_{edges}). The resulting quantity is shown in the graph at each time window.

Pearson-Partial and Kendall-Partial reveal trends remarkably different from the others. In particular two major drops appear in 2000 and 2006, with no evident connection to crises or important events. Yet they display a strong sensitivity to financial crises as well, in particular to the Russian crisis, Lehman Brothers default and Euro debt crisis (the latter being observed only in the Pearson-Partial edge overlap).

To complete the picture provided by the edge overlap, we have analysed the quantity $U^{[\alpha]}$ described in Eq. 7.3. In Fig. 7.4 we show U^α for each layer. As expected, Pearson and Kendall have the lowest values of U^α, since they are quite similar to each other and therefore share many edges. The trends of U^α in each layer are consistent with what observed previously on the edge overlap: Pearson, Kendall and Tail evolutions are strongly correlated, resulting in similar patterns of pairwise edge overlap in Fig. 7.3, whereas Partial U^α values fluctuate more independently. Nonetheless in the last three years the Partial layer seem to correlate more and more, especially with the Tail layer.

7.5.2 A multiplex cartography of network filtering

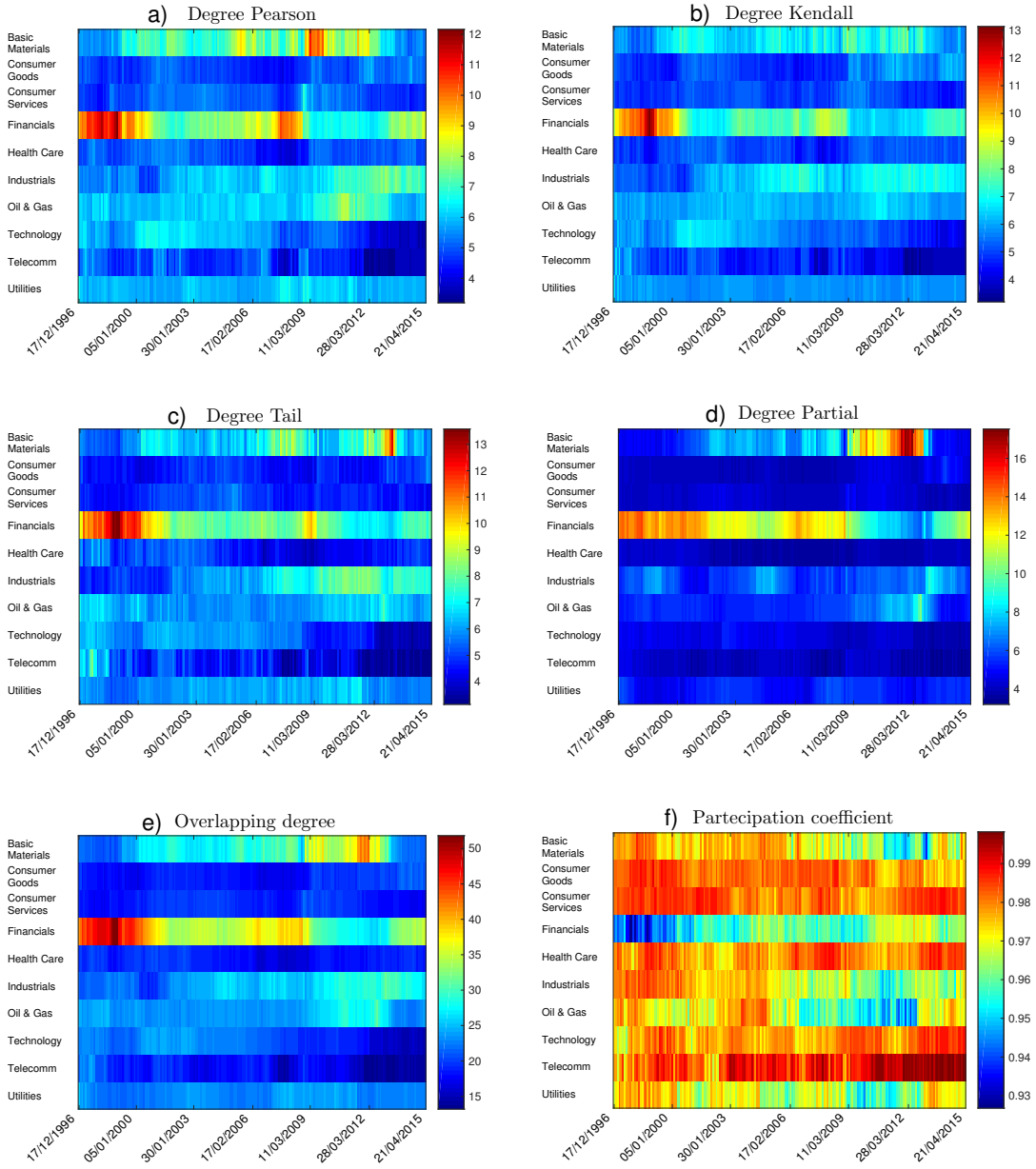

Fig. 7.5 **Comparison among degree evolution on different layers**. The average degree within each industry I on layer α, k_I^α, is shown for layer Pearson in a), for layer Kendall in b), for layer Tail in c) and layer Partial in d) (respectively, $\alpha = 1,2,3,4$), at each time window. In e) we show at each time window the average overlapping degree within each industry I, o_I. Finally in f) the average partecipation coefficient within each industry I, p_I, is shown for each time window.

Edge overlap and multidegree focus on the amount of information shared by layers in terms of common edges. Overlapping degree and partecipation coefficient, defined in

Eqs. 7.5 and 7.6, measure instead to what extent different layers agree on the centrality of each node, in terms of number of connections. In this section we focus on this aspect.

We have first calculated the degree evolution for each layer α separately, averaged over each ICB industry: $k_I^\alpha = \langle k_i^\alpha \rangle_{i \in I}$, where k_i^α is the degree of node i on layer α. In Fig. 7.5 a)-d) we show k_I^α at each time window for $\alpha = 1,2,3,4$ respectively. All layers assign to Financials the highest average degree, that has reached its peak in the late 90s before the Dot-com bubble and during the 2007-08 crisis. After that the average degree of Financials has dropped sensitively, but has started to recover in 2014. A part from these similarities, the picture is quite heterogeneous among layers. In Pearson layer, Basic Materials arises as second most central industry throughout most of the period, whereas Industrials and Oil & Gas acquired more connections in the post-crisis period in 2009. In Kendall layer the degree of all industries appears much attenuated, revealing a more homogeneous distribution of edges among nodes. Interestingly, Tail layer seems more similar to Pearson in this respect. Finally, Partial layer shows the highest level of concentration of links in Finance (consistently to what found in [71]) and, after the 2007-08 crisis, in Basic Materials.

The overlapping degree aggregates this information in a single quantity: we have calculated for each industry I the average overlapping degree, defined as follows:

$$o_I \equiv \langle o_i \rangle_{i \in I} \quad , \tag{7.21}$$

where o_i is the overlapping degree of node i. The dynamic result is shown in Fig. 7.5 e). As we can see o_I is able to highlight the preminence of Financials, Basic Materials, Oil & Gas and Industrials across the layers, providing at the same time a clearer and cleaner picture of degrees evolution. In particular four phases appear, clearly distinct from each others: the first, in which Financials is the only preminent industry, ends at the beginning of 2000; the second one lasts until the 2007-08 crisis and is characterized by the emergence of Basic Materials as second central industry; the third phase starts in 2009 and sees Financials loosing its preminence, in favour of Industrials, Oil & Gas

and Basic Materials (that becomes the most central); finally in 2014 a new equilibrium began, with Financials gaining again centrality followed by Industrials.

Finally, the partecipation coefficient completes the overlapping degree information, measuring the level of degree homogeneity among layers. Again, we have calculated an average within each ICB industry I, in formula:

$$p_I \equiv \langle p_i \rangle_{i \in I} \ , \tag{7.22}$$

where p_i is the partecipation coefficient of node i. Interestingly, as one can see from Fig. 7.5 f), p_I reveals that the main drops of homogeneity occur in Financials, Basic Materials, Industrials and Oil & Gas in correspondence to their main increase in overlapping degree. This fact indicates that increases in multiplex centrality (as measured by the overlapping degree) are mostly due to edges concentration on only a subset of layers (possibly one), consistently to what observed in the multidegree analysis.

These observations are further supported by Figs. 7.6 and 7.7, where for each industry I we have plotted the average partecipation coefficient p_I (x-axis) against the average overlapping degree o_I (y-axis). Each circle represents a different time window. Variations in circles size and colour mark different time windows. We can see how the two quantities appear anticorrelated (especially in Consumer Goods, Consumer Services, Industrials, Oil & Gas and Utilities) or at most uncorrelated. This again is a strong indication of the importance of monitoring all layers together, as an increase in the structural role of an industry (as measured by the overlapping degree) is tipically due to only a subset of layers (as indicated by the corresponding decrease of partecipation coefficient).

7.5.3 Identifying each contribution: multidegree

We can explore deeper single nodes' role across the layers by means of the concept of multidegree $k_i^{\vec{m}}$, defined in Eq. 7.8. We have computed multidegree for each node i

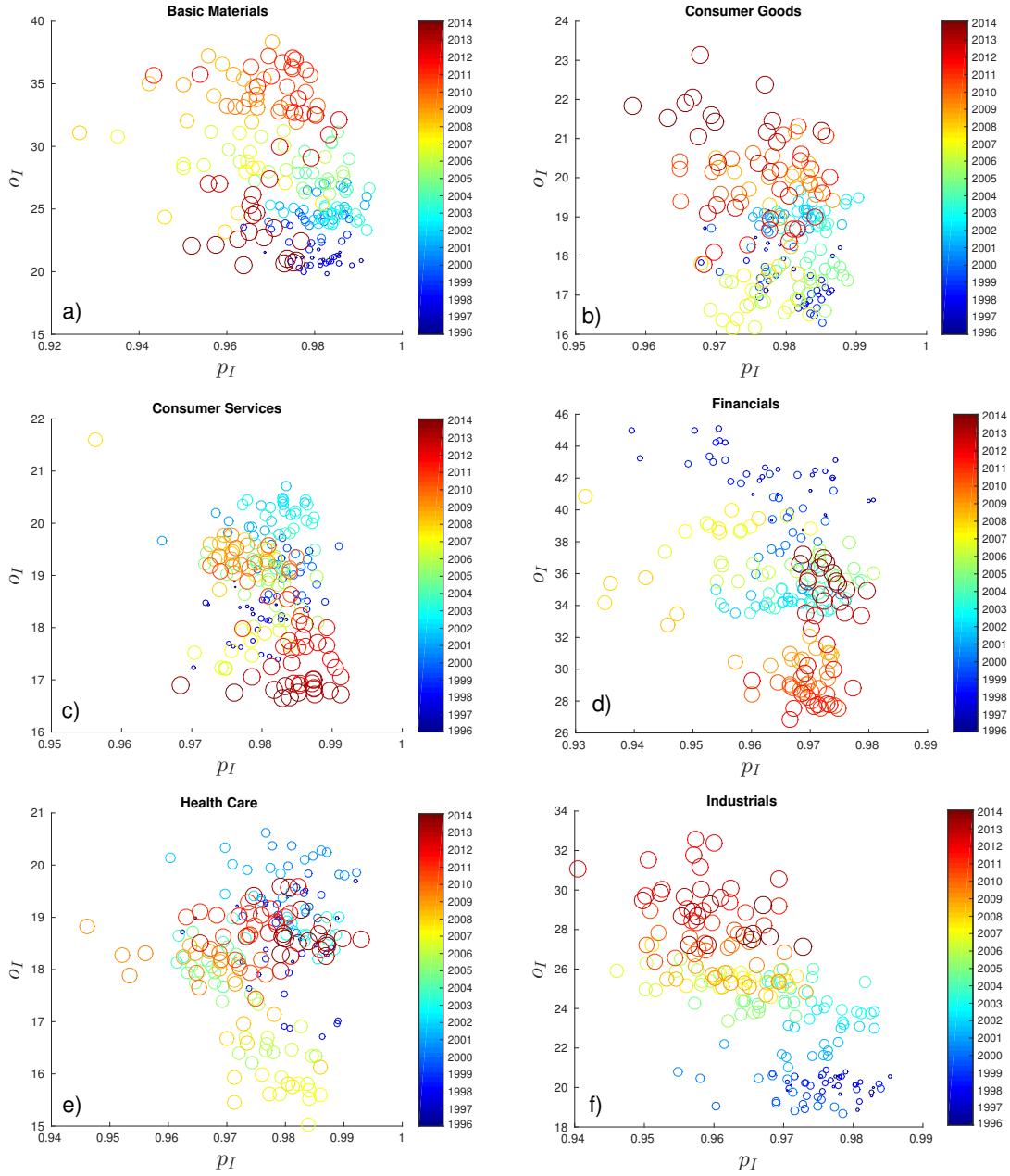

Fig. 7.6 **Industries evolution in the overlapping degree/partecipation coefficient plane (part 1)**. Fixed an industry I, we have plotted for each time window a circle whose y coordinate is the average overlapping degree o_I and whose x coordinate is the average partecipation coefficient p_I. Points at different times are characterized with different sizes (small to large) and colours (legend on the right). In a), b), c), d), e) and f) we show the results respectively for Basic Materials, Consumer Goods, Consumer Services, Financials, Health Care and Industrials.

at each time window. Since we are interested in the node contribution on each layer in terms of degree, we have normalised each node multidegree by the corresponding node overlapping degree o_i. The resulting $k_i^{\vec{m}}/o_i$ is the fraction of multiplex edges of

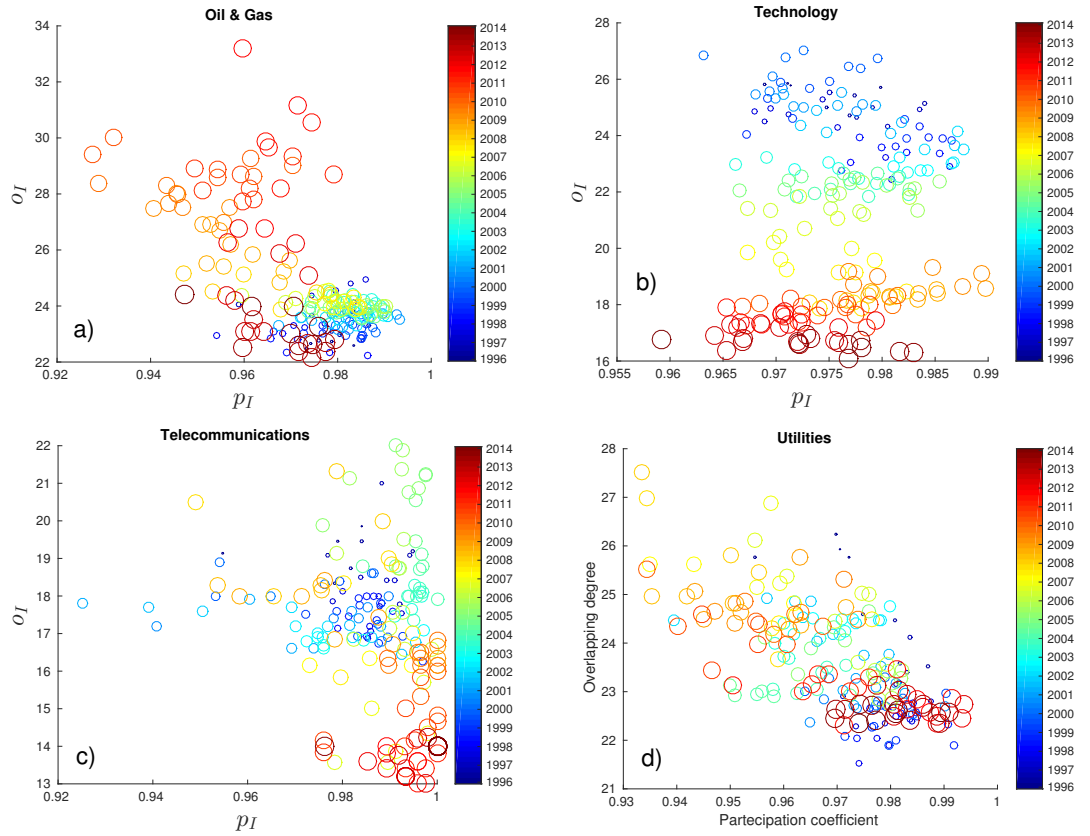

Fig. 7.7 **Industries evolution in the overlapping degree/partecipation coefficient plane (part 2)**. Fixed an industry I, we have plotted for each time window a circle whose y coordinate is the average overlapping degree o_I and whose x coordinate is the average partecipation coefficient p_I. Points at different times are characterized with different sizes (small to large) and colours (legend on the right). In a), b), c) and d) we show the results respectively for Oil & Gas, Technology, Telecommunications and Utilities.

node i that exist only on the subset of layers specified by $\vec{m} = (m_1, m_2, m_3, m_4)$. Finally we have averaged such quantity over all nodes belonging to the same ICB industry I: $\kappa_I^{\vec{m}} \equiv \langle k_i^{\vec{m}} / o_i \rangle_{i \in I}$. We will refer to this quantity as the normalised multidegree of industry I.

In Fig. 7.8 this quantity is shown in time, for different choiches of vector $(\vec{m})$. We focus, among the $2^4 = 16$ possible multidegrees, on those that correspond to the number of edges that each node i has only on layer Pearson ($k_i^{(0,0,0,1)}$), Kendall ($k_i^{(0,0,1,0)}$), Tail ($k_i^{(0,1,0,0)}$), Partial ($k_i^{(1,0,0,0)}$), and on at least one among Kendall, Tail and Partial, but not on Pearson ($\sum_{m_1,m_2,m_3=0,1} k_i^{(m_1,m_2,m_3,0)}$). As one can see there is high heterogeneity:

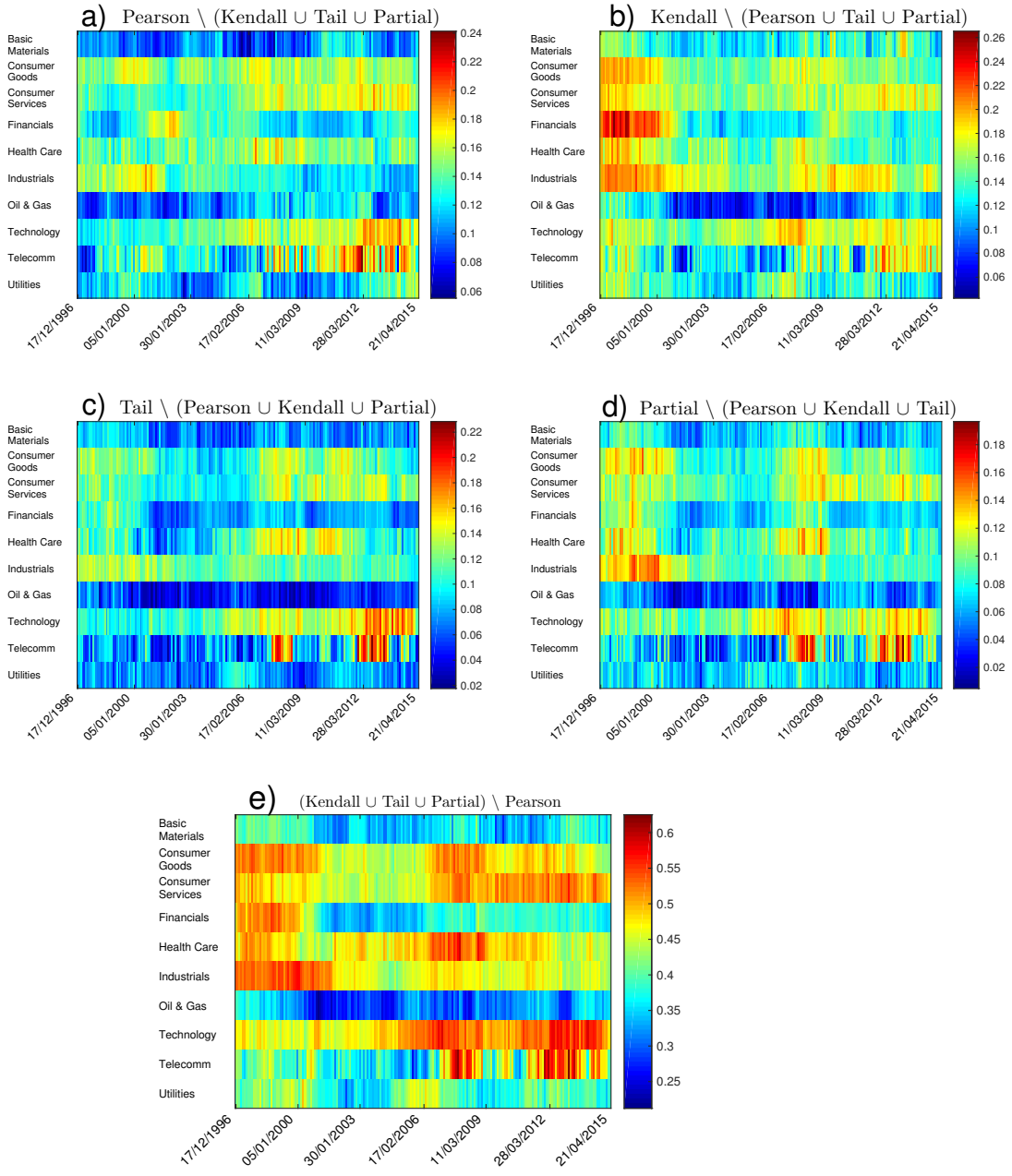

Fig. 7.8 **Normalised multidegree for each ICB industry I, $\kappa_I^{\vec{m}}$, at different times**. To obtain normalised multidegree the multidegree $k_i^{\vec{m}}$ of each node i is first normalised by its overlapping degree o_i and then averaged over all nodes belonging to the industry I (see text for further details). a) $\kappa_I^{(0,0,0,1)}$, corresponding to edges existing only on Pearson layer; b) $\kappa_I^{(0,0,1,0)}$, corresponding to edges existing only on Kendall layer; c) $\kappa_I^{(0,1,0,0)}$, corresponding to edges existing only on Tail layer; d) $\kappa_I^{(1,0,0,0)}$, corresponding to edges existing only on Partial layer; e) $\sum_{m_1,m_2,m_3=0,1} \kappa_I^{(m_1,m_2,m_3,0)}$, corresponding to edges existing on at least one layer among Kendall, Tail and Partial, but not on Pearson.

there are some industries - such as Oil & Gas, Utilities and Basic Materials - that have relatively low values of normalised multidegree, pointing out a strong agreement among the different layers in assessing their significant connections in the market; and others - Industrials, Finance, Technology, Telecommunications and Consumer Services - whose edges tend to concentrate on one or a subset of layers. In particular it is worth noting that, in the years preceeding the Dot-com bubble and 2002 downturn, there was a higher concentration of Finance, Industrials and Consumer Goods edges on Kendall layer only, whereas in the post 2007-08 crisis a sudden increase of edges existing only on Tail occured for Consumer Goods, Consumer Services and Health Care. These observations indicate the importance of using more than one measure of dependence to describe completely the market structure.

From this perspective it is worth discussing the normalised multidegree corresponding to the union of Kendall, Tail and Partial without Pearson in Fig. 7.8: this picture gives information on those connections in the market that are detected by all layers but Pearson. As one can see, until 2002 a Pearson analysis would have missed from 40% up to 60% of edges in industries such as Basic Materials, Financial, Consumer Goods and Industrials. After 2002 these percentages tend to decrease until the 2007-08 crisis, when high concentrations missed by Pearson can be observed especially in Consumer and Health Care industries. The period following the 2007-08 crisis is also characterized by a sensitive and unprecedented rise of normalised multidegree from Technology and Telecommunications, whose importance in the market dependence structure has been therefore underestimated by Pearson over the last ten years.

7.5.4 Interlayer degree-degree correlation: a comparison of assets centrality ranking

In order to analyse further the heterogeneity of degree, we have calculated the interlayer degree-degree correlation that allows us to compare pairs of layers separately. Namely, for each pair of layers and each time window we have calculated the quantity $\rho^{deg}_{\alpha_1\alpha_2}$

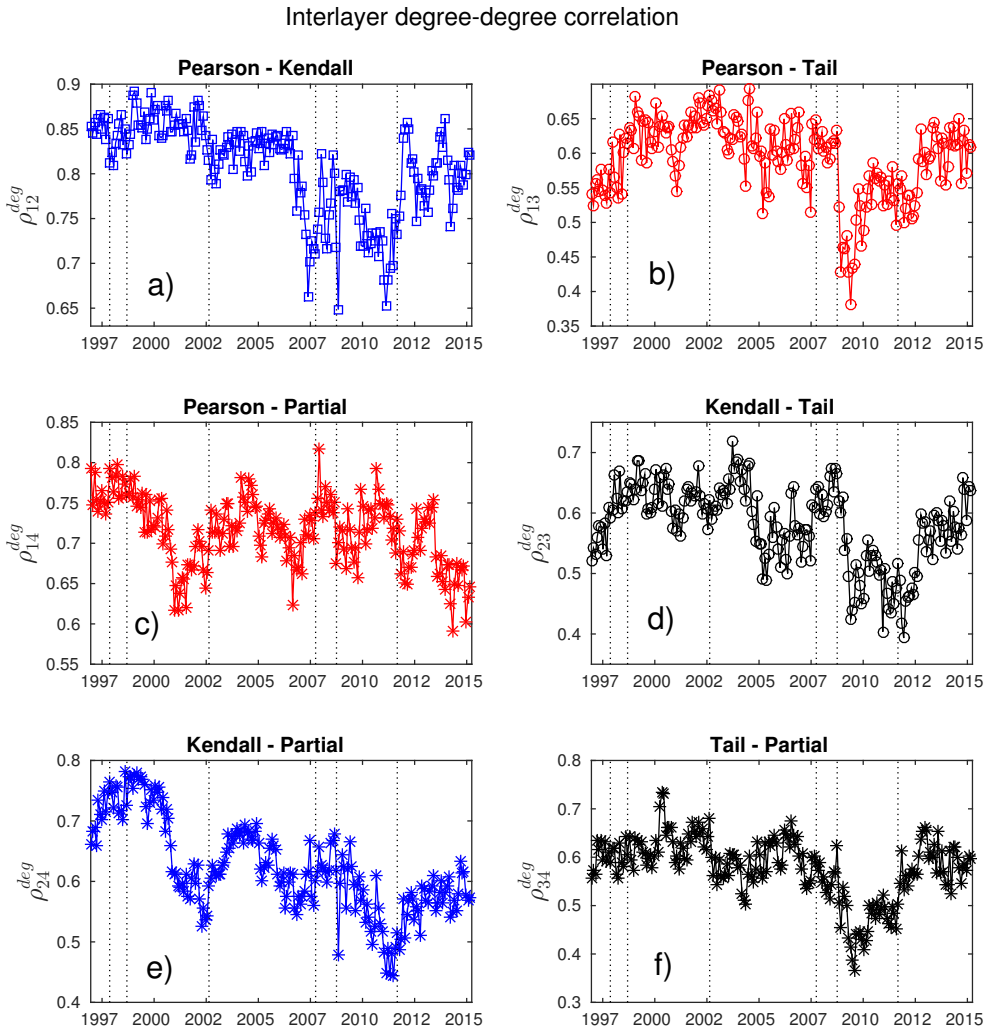

Fig. 7.9 **Interlayer correlation for each pair of layers, at different time windows**. The graphs show the interlayer degree-degree correlation $\rho^{deg}_{\alpha_1\alpha_2}$ between Pearson and Kendall in a), Pearson and Tail in b), Pearson and Partial in c), Kendall and Tail in d), Kendall and Partial in e) and Tail and Partial in f).

described in Eq. 7.4. The graphs we have obtained are in Fig. 7.9. The curves $\rho^{deg}_{\alpha_1\alpha_2}$ for each pair are only weakly correlated with the corresponding edge overlaps $\langle O \rangle_{\alpha_1\alpha_2}$ in Fig. 7.3. This is not surprising, as they are carrying a different and complementary information, namely the similarity between the rankings of degrees in the two layers. The main change in the interlayer correlation is due to the 2007-08 crisis, that triggers a drop of degree correlation. Pre-crisis levels are reached again only after 2012. It is

worth noticing that the correlation between Pearson and Kendall starts decreasing in 2006 and hits a minimum few months before the 2007 turbulent period: this pre-crisis trend was evident in the edge overlap as well (Fig. 7.3). As in the edge overlap analysis, the pairs Pearson-Partial and Kendall-Partial display features very different from the others, with quite a volatile evolution that does not seem to be connected to the main events affecting the financial market.

7.6 Summary

In this chapter we have exploited the versatility of multiplex metrics to investigate the degree of non-linearity in the global dependence structure of equity returns. To this aim, we have constructed four different layers of PMFG by using as many dependence measures, and analysed similarities and differences among their topologies in a rolling time window setting.

According to the mean edge overlap between the first three layers, the importance of non-linearity and tails on market dependence structure has dropped significantly in the first half of 2000s. Then it has risen steeply between 2005 and the 2007-08 crisis, and ever since is slowly diminishing. The analysis of mean edge overlap with the fourth layer reveals instead that the role of partial correlation in the dependence structure became increasingly important in 2000, 2006, 2012 and 2014. Overall, financial crises triggered remarkable drops in the edge overlap, widening therefore the differences among the measures of dependence just when evaluation of risk becomes of the highest importance.

The multidegree analysis reveals that different industries exhibit different levels and patterns of edge overlap in time. In particular, Financials, Industrials and Consumer Goods show an increasing number of connections only on Kendall layer in the late 90s/early 2000, at the edge of the Dot-Com bubble; overall these industries tend to have many edges on layers Kendall, Tail and Partial that Pearson does not contain, especially

after the 2007-08 crisis: this fact points out further the importance of not relying only on Pearson estimator for dependence analysis.

Significant changes in assets degree centrality tend to occur on one layer (or a subset of layers) only, without sensitive correlation among different measures of dependence: this is revealed by the analyses on overlapping degree and partecipation coefficient. Nonetheless, the extent to which different layers agree on ranking assets centrality is a highly dynamic quantity, that the interlayer degree-degree correlation reveals to have dropped dramatically after the 2007-08 financial crisis. Assets centrality in a dependence network is an important feature for portfolio optimization and risk management [80, 81, 78]: our findings highlight the importance of monitoring these features by means of more than one measure of dependence.

Chapter 8

Conclusions and Outlook

In this thesis we have studied the complexity underlying the dependence structure of financial time series. We have relied on the network filtering techniques to tackle problems such as non-stationarity and non-linearity of financial correlation. We have mainly unveiled novel empirical properties that had never been detected before, and proposed a method to take advantage of them for forecasting market volatility.

Specifically, we have applied the DBHT method to financial data for the first time in Chapter 4, where we have compared its performance in terms of economic information and stability with four other clustering methods. We have found that such features depend heavily on the underlying algorithm of each method, suggesting that the application of clustering to asset allocation should take into account such differences. We have then focused on the temporal evolution of dependence. The dynamical analysis of DBHT clustering has shown that the market mode factor has become more influential over the last 15 years, revealing a pattern which was not observed without clustering. This trend began well before the financial crisis and seems to have a much longer time horizon. In Chapter 5, we have demonstrated how the financial crisis can be viewed as a phase transition between two structurally different dependence structures. Notably, in the post-crisis phase the industrial classification is less relevant for risk diversification. Moreover, we have reported for the first time evidence of long-term memory in the

evolution of correlation-based networks, which opens interesting scenarios for the modelling of dependence structure evolution. An explicit connection between filtered networks dynamics and market risk is then unveiled in Chapter 6, where we have reported a relation between rate of change in the dependence structure and variations of market volatility. We have proposed to use these empirical facts to predict future changes in volatility. Such a forecasting tool outperforms methods based on past volatility only. Besides, it overcomes the curse of dimensionality, which limits the applicability of many traditional econometrics techniques, therefore making it valuable for systemic risk and early-warning analyses. Finally, in Chapter 7 we have relied on network filtering to investigate non-linear dependencies among asset returns from a global perspective. To this end, we have combined for the first time the multiplex framework with network filtering. Through this approach, we have revealed that networks built from Pearson, Kendall, Tail and Partial correlation display deep differences, which also change in time. Hence, the influence of non-linearity on the dependence structure is changeable, making the use of Pearson coefficient alone not reliable for portfolio optimisation and risk management, especially during financial crises. In particular, different industrial sectors display different patterns and degrees of non-linearity. To the best of our knowledge this is the first analysis investigating the issue of non-linear dependence from a market level perspective.

To summarise, main results were the following:

- We have applied, for the first time, the DBHT method to financial data, highlighting its advantages over the other hierarchical clustering techniques.

- We have quantified and compared the economic information extracted from five clustering methods, revealing how they yield quite different performances and stability. We have discussed the implication of these differences for applications to portfolio optimisation.

- We have studied the dynamical evolution of the DBHT clustering tracking single clusters, the whole community partition and the correlation-based network. Following this method, we have found strong evidence of non-stationarity during the financial crisis, as well as a phase transition which made the post-crisis period structurally different from the pre-crisis one.

- We have proposed a new method which uses network filtering to predict future changes in market volatility. This new tool overcomes the limits of traditional econometric techniques when it comes to dealing with hundreds of assets.

- We have applied for the first time multiplex metrics to correlation-based networks, tackling the problem of non-linearity from a global perspective. This original approach has revealed that non-linearity affects the global dependence structure in a deep and changeable way, which also depends on the industrial sector. In particular rankings of assets centrality are strongly dependent on the dependence measure which is used.

The analyses carried out in this thesis are very promising in opening new questions and research ideas. Non-stationarity and non-linearity are two challenging aspects of financial correlation which still need to be satisfactorily addressed and incorporated in Risk Management and Portfolio Optimisation. In this respect, the use of network filtering that we have proposed in this thesis may well contribute to make traditional techniques better at dealing with changing market scenarios, when the traditional assumptions of stationarity and linearity fail. The results we have presented suggest different ways this contribution can be achieved. We have shown that cluster tracking reveals the build-up of a financial cluster in the months preceding the financial crisis; further research should be carried out to translate such patterns in reliable early-warning signals. Generally, understanding how to monitor the stability of the clustering structure would provide an intuitive and powerful tool for detecting the emergence of new risk

factors. From the methodological point of view, valuable insights might come from the research strand on anomaly detection problems [220].

From a more theoretical perspective, our findings on long-term memory in the correlation-based networks evolution are promising as well. The existence of patterns in the dynamics of networks topology indicates that further effort should be made to model such evolution, similarly to what has been done in other areas of Network Theory [221]. Understanding and anticipating correlation-based network changes could prove to be crucial for predicting structural shift in the underlying dependence structure.

Another interesting finding of this thesis is the aforementioned increase of the market mode influence over the last 15 years, as demonstrated by the dynamical comparison between detrended and non-detrended DBHT clusterings. Interestingly this pattern is visible only through the clustering analysis, since no steady trend is visible from the evolution of the average correlation. The average correlation is to a very good approximation equal to the first principal component of the market, as defined in the Principal Component Analysis [33]; therefore this result suggests that there are patterns in the dependence evolution which cannot be revealed by using spectral methods. In this sense, a deeper integration of clustering and network filtering into factor models would be advantageous, so that they can be effectively exploited in all their applications (from Pricing to Asset allocation) where PCA is currently the most popular dimensionality reduction technique for choosing the factors. A promising result in this direction is presented in [222], where the authors show how to construct a statistically robust factor model from Linkage dendrograms. Extensions of this model to take into account the dynamical aspects of clustering discussed in this thesis would be of great interest.

As far as the prediction of volatility is concerned, our findings represent a first step towards the use of network filtering in Econometrics. The performance and flexibility of our forecasting tool should be checked across different asset classes, from currencies to fixed-income assets. A further step could be the application to large portfolios made of mixed asset classes, taking advantage of the scalability offered by the recently

introduced TMFG filtering [181]. Since our tool is suitable for portfolios made of several assets whereas multiGARCH models cope with few assets, the two approaches are complementary and it would be interesting to combine them to achieve more accuracy in the forecasting. The empirical relation between dependence structure persistence and volatility, which we have reported and exploited for the forecasting tool, is worth investigating further. We have interpreted such relation in terms of dynamics of risk factors in the market. Further analyses should be devoted to confirm this interpretation and understand its implication for the application of network filtering to factor models.

The multiplex analysis we have presented represents a first step towards a proper research strand on correlation-based multiplex networks. Given the diversity of dependence measures that are available, multiplex network filtering is the natural prosecution of the research carried out over the last 15 years on single-layer correlation-based networks. The analysis we have presented here could be performed by using other dependence measures, such as those based on information theory [223, 206, 207], as well as applying different multiplex metrics. The implications for the applications are likewise intriguing. We have shown how multiplex metrics can quantify the degree of non-linearity in the dependence structure; one should investigate how to translate these measures into information valuable for portfolio optimisation techniques. A possible approach could be using multiplex metrics as indicators that signal which dependence measure is more suitable for the current market scenario.

Appendix A

DBHT algorithm

The general idea at the basis of the DBHT method [66, 83] is to exploit the topological structure of PMFG graphs to construct a hierarchical clustering over the set of nodes. The PMFG is assumed to be a weighted graph, whose edge-weights represent measures of similarity among nodes (correlation if the PMFG is computed from correlation matrix). Moreover a dissimilarity measure is assumed to be associated to each edge as well.

As a direct consequence of planarity, any cycle (that is a closed path in the graph, with same starting and ending node) in G must be either separating or non-separating [224]. A separating cycle is a cycle which makes the graph disconnect into two disjoint and non-empty graphs if detached [66]. The idea behind the DBHT method is to use this property to identify a natural hierarchy from the PMFG topology. The simplest cycle is the 3-clique, which is also the key atomic structure of a PMFG [83]. Let us denote a separating 3-clique with k_p, and the two disconnected subgraphs connected by k_p with G_p^{ex} and G_p^{in}.

The union of k_p with either G_p^{ex} or G_p^{in} are still planar. We call these planar subgraphs "bubbles". After identifying all separating 3-cliques and corresponding bubbles, we can draw a diagram where bubbles are vertex connected by edges which represent the corresponding 3-cliques. It turns out that such diagram is always a tree, called "bubble

tree" [66]. This tree is the backbone of the DBHT hierarchical structure. Let us call $\{b_i\}_i$ the set of bubbles in the bubble tree, and let $b_i b_j$ the separating 3-clique connecting bubbles b_i and b_j. The rest of the algorithm aims to infer a proper hierarchical clustering from this structure [66]:

1. **Identifying converging bubbles**: to each separating clique $b_i b_j$ in the bubble tree can be associated a direction towards either b_i or b_j, depending on which bubble the clique is connected to with more strength. The strength is computed summing weights over all edges connecting $b_i b_j$ with b_i and b_j [66]:

$$W_p^{i/j} = \sum_{v \in k_p, u \in b_{i/j}} G(u,v) \ , \tag{A.1}$$

   where $G(u,v)$ is the u,v entry of the PMFG adjacency matrix, equal to the similarity between nodes u and v. Once a direction is defined for all the separating 3-cliques, we can distinguish between converging bubbles, where the connected edges are all incoming to the bubble, diverging bubbles, where the connected edges are all outcoming from the bubble, and passage bubbles, where there are both incoming and outcoming connected edges. To each converging bubble we associate a cluster. Nodes which belong to the converging bubble are assigned to the corresponding cluster. Moreover, nodes which are in bubbles that are connected by a directed path to a converging bubble b_α belong to the cluster α. We denote the subtree of bubbles connected to a converging bubble α with $\vec{h_\alpha}$.

2. **Defining a discrete clustering**: following the previous step some nodes could be assigned to more than one cluster, ending up with a non-discrete clustering. Indeed, some nodes might belong to more than one converging bubble. In this case, we look at the node strength of attachment to each bubble [66]:

$$\chi(v, b_\alpha) = \frac{\sum_{v \in V(b_\alpha)} G(u,v)}{3(|V(b_\alpha)| - 2)} \ , \tag{A.2}$$

where $|V(b_\alpha)|$ is the number of nodes in the bubble b_α and $3(|V(b_\alpha)|-2)$ is the number of edges in the bubble. We then assign the node to the bubble/cluster which shows the largest strength. Let us call $V^0(\alpha)$ the set of nodes belonging to cluster α after this procedure has been performed. Another scenario that leads to non-discrete clustering is when a node belong to a non-converging bubble which is connected to more than one converging bubble. In this case we assign the node to the bubble to which the node is closer in terms of average shortest path distance [66]:

$$\bar{L}(v,\alpha) = mean\{l(v,u)|u \in V^0(\alpha) \wedge V(\vec{h_\alpha})\} \ , \tag{A.3}$$

where $l(v,u)$ is the shortest distance between nodes u and v (namely the smallest sum of dissimilarities $D(r,s)$ over any path between u and v), and $V(\vec{h_\alpha})$ is the set of nodes in the subtree $\vec{h_\alpha}$. Eventually we assign to each node a unique cluster. We denote with $V(\alpha)$, $V(\beta)$ ... the set of nodes in clusters α, β ...

3. **Intra-cluster hierarchy**: we can build a hierarchy among nodes within each clusters in two steps. We first construct a hierarchy inside each bubble. This is achieved by performing a Complete Linkage by using the shortest distance $l(u,v)$ as distance matrix. Nodes which belong to more than one bubble are assigned to the bubble which has the highest strength of attachment $\chi(v,b_\alpha)$. Then we perform a Complete Linkage among the bubbles in each cluster α, by using the distance matrix [66]:

$$d^I_\alpha(b_i,b_j) = \max\{l(u,v)|u \in V^\alpha(b_i) \wedge v \in V^\alpha(b_i)\} \ , \tag{A.4}$$

where $V^\alpha(b_i/b_j)$ is the set of nodes belonging to bubble b_i/b_j in cluster α. In this way we obtain a hierarchical order among nodes inside each cluster.

4. **Inter-cluster hierarchy**: finally, a Complete Linkage procedure is performed among clusters to construct a inter-cluster hierarchy. The distance matrix used is [66]:

$$d^{II}(\alpha,\beta) = \max\{l(u,v)|u \in V(\alpha) \wedge v \in V(\beta)\} \quad . \tag{A.5}$$

Once this last procedure is carried out, we obtain a hierarchical order among clusters. The dendrogram structure is therefore complete: starting from the discrete clustering obtained at step 2, we have a hierarchy both within each cluster and among different clusters.

Appendix B

Bootstrapping

In statistics, bootstrapping is a tool that relies on repeatedly drawing samples from a given dataset, in order to perform statistical tests or error estimations [190, 14]. It belongs to the broader class of resampling techniques [14]. In particular, bootstrapping uses random sampling with replacement to construct a family of approximated distributions of the original population. Then the parameter of interested is estimated from each bootstrapped sample. This collection of estimated parameters can finally be used to compute both the error and the confidence interval for the parameter. It is often used as an alternative to parametric statistical test, as it does not require any assumption about the true distribution.

Let us here describe in more details the bootstrapping procedure [190]. Let us call X a random variable (or a vector of random variables), with (unknown) distribution F. If X is a vector, F is a multivariate distribution. We are interested in an estimator $T(X)$ and in its statistical properties, such as variance or its confidence interval. Given a sample $x = \{x_1, x_2, ..., x_n\}$ from F, we can compute the estimator by using this sample, namely $T(x)$. However, in order to estimate its statistical properties we need information about F, which is not available. Bootstrapping overcomes this problem by resampling with replacement from $x_1, x_2, ..., x_n$, providing a bootstrapped sample $x^* = x_1^*, x_2^*, ..., x_n^*$ and a corresponding replica estimator $T(x^*)$. By repeating the resampling B times, we

obtain B independent bootstrap replicas $T(x^{*1}), T(x^{*1}), ..., T(x^{*B})$. From these family of bootstrap replicas we can compute the standard deviation of $T(x)$ as follows [190]:

$$SD_T = \sqrt{\frac{1}{B-1}\sum_{r=1}^{B}\left(T(x^{*r}) - \frac{1}{B}\sum_{r'=1}^{B}T(x^{*r'})\right)^2}. \tag{B.1}$$

Confidence intervals for the estimator can be constructed as follows [191]. Let us compute for each replica the quantity $\delta^{*i} = T(x^{*i}) - T(x)$. Let us call $\delta^*_{\alpha/2}$ and $\delta^*_{1-\alpha/2}$ respectively the $\alpha/2$ and $1-\alpha/2$ percentiles of this population. Then the α confidence interval for $T(X)$ is [191]:

$$\left(T(x) - \delta^*_{\alpha/2} \;;\; T(x) - \delta^*_{1-\alpha/2}\right) \;. \tag{B.2}$$

If the observations are dependent, this approach is not suitable because resampling with replacement fails to replicate this dependence. This scenario can occur for example with autocorrelated time series. Block-bootstrapping [199] overcomes this problem by resampling blocks of data instead of single observations. Blocks can be either non-overlapping or overlapping [199]. As for the optimal block length, different approaches have been proposed which take into account the length of autcorrelation [203, 225].

References

[1] G. Parisi. Complex systems: a physicist's viewpoint. *Physica A*, 263:557–564, 1999.

[2] M. Scheffer, J. Bascompte, W. A. Brock, V. Brovkin, S. R. Carpenter, V. Dakos, H. Held, E. H. van Nes, M. Rietkerk, and G. Sugihara. Early-warning signals for critical transitions. *Nature*, 461:53–59, doi:10.1038/nature08227, 2009.

[3] R. N. Mantegna and H. E. Stanley. *An Introduction to Econophysics*. Cambridge University Press, 2000.

[4] M.M. Dacorogna. *An introduction to high-frequency finance*. Academic Press, 2001.

[5] A. Chakraborti, I.M. Toke, M. Patriarca, and F. Abergel. Econophysics review: I. empirical facts. *Quantitative Finance*, 11:991–1012, 2011.

[6] J.-P. Bouchaud and M. Potters. *Theory of Financial Risk and Derivative Pricing: From Statistical Physics to Risk Management*. Cambridge, 2000.

[7] J.-P. Bouchaud. Crises and collective socio-economic phenomena: Simple models and challenges. *J. Stat. Phys.*, 151:567–606, DOI 10.1007/s10955–012–0687–3, 2013.

[8] D. Sornette. *Why Stock Markets Crash: Critical Events in Complex Financial Systems*. Princeton University Press, 2004.

[9] W. Feller. *An Introduction to Probability Theory and Its Applications*. John Wiley, 1968.

[10] O. Kallenberg. *Foundations of Modern Probability*. 2nd edition, Springer Series in Statistics, 2002.

[11] D. B. West. *Introduction to graph theory*. Prentice-Hall, Englewood Cliffs NJ, 1996.

[12] S. N. Dorogovtsev. *Lectures on Complex Networks*. Oxford, 2010.

[13] G. Caldarelli. *Scale-Free Networks*. Oxford Finance Series, 2007.

[14] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer, 2014.

[15] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2007.

[16] G. Bonanno, F. Lillo, and R.N. Mantegna. Levels of complexity in financial markets. *Physica A*, 299 (1):16–27, 2001.

[17] B. Mandelbrot. The variation of certain speculative prices. *The Journal of Business*, 36 (4):394–419, 1963.

[18] R. J. Buonocore, N. Musmeci, T. Aste, and T. Di Matteo. Two different flavours of complexity in financial data. *EPJ ST, submitted*, 2016.

[19] V. Plerou, P. Gopikrishnan, L. A. Nunes Amaral, M. Meyer, and H. E. Stanley. Scaling of the distribution of price fluctuations of individual companies. *Phys. Rev. E*, 60 (6):6519 – 6529, 1999.

[20] P. Gopikrishnan, V. Plerou, L.A. Nunes Amaral, M. Meyer, and H.E. Stanley. Scaling of the distribution of fluctuations of financial market indices. *Phys. Rev. E*, 60 (5):5305–5316, 1999.

[21] Di Matteo T. Multi-scaling in finance. *Quantitative finance*, 7 (1):21–36, 2007.

[22] J. Barunik, T. Aste, T. Di Matteo, and R. Liu. Understanding the source of multifractality in financial markets. *Physica A*, 391:4234–4251, 2012.

[23] Di Matteo T., T. Aste, and M.M. Dacorogna. Scaling behaviors in differently developed markets. *Physica A*, 324:183–188, 2003.

[24] T. Di Matteo, T. Aste, and M. M. Dacorogna. Long term memories of developed and emerging markets: using the scaling analysis to characterize their stage of development. *Journal of Banking & Finance*, 29/4:827–851, 2005.

[25] R. Morales, T. Di Matteo, and T. Aste. Non-stationary multifractality in stock returns. *Physica A*, 392:6470–6483, 2013.

[26] R. Morales, T. Di Matteo, and T. Aste. Dependency structure and scaling properties of financial time series are related. *Sci. Rep.*, 4:4589, 2014.

[27] R. Morales, T. Di Matteo, R. Gramatica, and T. Aste. Dynamical generalized hurst exponent as a tool to monitor unstable periods in financial time series. *Physica A*, 391:3180–3189, 2012.

[28] R.J. Buonocore, T. Aste, and Di Matteo T. Measuring multiscaling in financial time series. *Chaos, Solitons and Fractals*, 2015.

[29] N. Nava, T. Di Matteo, and T. Aste. Anomalous volatility scaling in high frequency financial data. *Physica A*, 447:434–445, 2016.

[30] R. S. Tsay. *Analysis of financial time series*. John Wiley, 2005.

[31] J. W. Kantelhardt, S. A. Zschiegnera, E. Koscielny-Bundec, S. Havlind, A. Bundea, and H. E. Stanley. Multifractal detrended fuctuation analysis of nonstationary time series. *Physica A*, 316:87–114, 2002.

[32] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, T. Gühr, and H.E. Stanley. Random matrix approach to cross-correlations in financial data. *Phys. Rev. E*, 65:066126, 2002.

[33] L. Laloux, P. Cizeau, M. Potters, and J.-P. Bouchaud. Random matrix theory. *Risk*, 12:69, 1999.

[34] G. Bonanno, G. Caldarelli, F. Lillo, and R.N. Mantegna. Topology of correlation-based minimal spanning trees in real and model markets. *Phys Rev E Stat Nonlin Soft Matter Phys*, 68:046130, 2003.

[35] V. Plerou, P. Gopikrishnan, B. Rosenow, L.A.N. Amaral, and H.E. Stanley. Universal and nonuniversal properties of cross correlations in financial time series. *Phys. Rev. Letters*, 83:1471, 1999.

[36] L. Laloux, P. Cizeau, J.P. Bouchaud, and M. Potters. Noise dressing of financial correlation matrices. *Phys. Rev. Letters*, 83, 1999.

[37] M. Potters, J.-P. Bouchaud, and L. Laloux. Financial applications of random matrix theory: old laces and new pieces. *Acta Physica Polonica B*, 36, 2005.

[38] G. Livan, J. Inoue, and E. Scalas. On the non-stationarity of financial time series: impact on optimal portfolio selection. *J. Stat. Mech.*, page P07025, 2012.

[39] M.C. Munnix, T. Shimada, R. Schäfer, F. Leyvraz, T.H. Seligman, T. Guhr, and H.E. Stanley. Identifying states of a financial market. *Sci. Rep.*, 2:644, 2012.

[40] G. Raffaelli and M. Marsili. Dynamic instability in a phenomenological model of correlated assets. *J. Stat. Mech. E*, page L08001, 2006.

[41] M. Marsili, G. Raffaelli, and B. Ponsot. Dynamic instability in generic model of multi-assets markets. *J. Econ. Dyn. Control*, 33:1170, 2009.

[42] M. Raddant and F. Wagner. Phase transition in the s&p stock market. *J. Econ. Interact Coord.*, pages DOI 10.1007/s11403–015–0160–x, 2015.

[43] V. DeMiguel, L. Garlappi, and R. Uppal. Optimal versus naive diversification: how inefficient is the 1/n strategy? *Review of Financial Studies*, 22 (5):1915–1953, 2009.

[44] L. Bauwens, S. Laurent, and J.V.K. Rombouts. Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, 21:79–109, 2006.

[45] P.K. Clark. A subordinate stochastic process model with finite variance for speculative prices. *Econometrica*, 41:135–155, 1973.

[46] C. M. Hafner and H. Manner. Multivariate time series models for asset prices. *Handbook of Computational Finance*, pages 89–115, 2012.

[47] J.-P. Bouchaud and M. Potters. More stylized facts of financial markets: leverage effect and downside correlations. *Physica A*, 299:60–70, 2001.

[48] M. Andersson, E. Krylova, and S. Vähämaa. Why does the correlation between stock and bond returns vary over time? *Applied Financial Economics*, 18:139–151, 2008.

[49] F. Longin and B. Solnik. Extreme correlation of international equity markets. *The Journal of Finance*, 55:649–676, 2001.

[50] A. Anga and J. Chenb. Asymmetric correlations of equity portfolios. *Journal of Financial Economics*, 63:443–494, 2002.

[51] P. Embrechts, A. McNeil, and D. Straumann. Correlation and dependency in risk management: Properties and pitfalls. *Dempster, M. et al. (Eds.), Risk Management: Value at Risk and Beyond. Cambridge University Press, Cambridge*, 2001.

[52] D. Brigo, A. Pallavicini, and R. Torresetti. *Credit Models and the Crisis: A Journey Into CDOs, Copulas, Correlations and Dynamic Models*. Wiley, 2010.

[53] R. N. Mantegna. Hierarchical structure in financial markets. *Eur. Phys. J. B*, 11:193, 1999.

[54] J. P. Onnela, Anirban Chakraborti, Kimmo Kaski, Janos Kertész, and Antti Kanto. Asset trees and asset graphs in financial markets. *Phys. Scr.*, T106:48, 2003.

[55] T. Aste, T. Di Matteo, and S. T. Hyde. Complex networks on hyperbolic surfaces. *Physica A*, 346:20, 2005.

[56] M. Tumminello, T. Aste, T. Di Matteo, and R.N. Mantegna. A tool for filtering information in complex systems. *Proc. Natl. Acad. Sci.*, 102:10421–10426, 2005.

[57] T. Di Matteo and T. Aste. How does the eurodollars interest rate behave? *J. Theoret. Appl. Finance*, 5:122–127, 2002.

[58] T. Di Matteo, T. Aste, and R. N. Mantegna. An interest rate cluster analysis. *Physica A*, 339:181–188, 2004.

[59] T. Di Matteo, T. Aste, S. T. Hyde, and S. Ramsden. Interest rates hierarchical structure. *Physica A*, 335:21–33, 2005.

[60] M. Bartolozzi, C. Mellen, T. Di Matteo, and T. Aste. Multi-scale correlations in different futures markets. *Eur. Phys. J. B*, 58:207–220, 2007.

[61] T. Aste and T. Di Matteo. Dynamical networks from correlations. *Physica A*, 370:156–161, 2006.

[62] M. R. Anderberg. *Cluster analysis for applications*. Academic Press, 1973.

[63] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. New York: Springer, 2009.

[64] M. Tumminello, F. Lillo, and R. N. Mantegna. Correlation, hierarchies, and networks in financial markets. *J. Econ. Behav. Organ.*, 75:40–58, 2010.

[65] P. H. A. Sneath. The application of computers to taxonomy. *Journal of General Microbiology*, 17 (1):201–226, 1957.

[66] W. M. Song, T. Di Matteo, and T. Aste. Hierarchical information clustering by means of topologically embedded graphs. *PLoS ONE*, 7:e31929, 2012.

[67] M. Tumminello, T. Di Matteo, T. Aste, and R. N. Mantegna. Correlation based networks of equity returns sampled at different time horizons. *Eur. Phys. J. B*, 55:209–217, 2007.

[68] C. Borghesi, M. Marsili, and S. Micciché. Emergence of time-horizon invariant correlation structure in financial returns by subtraction of the market mode. *Phys. Rev. E*, 76:026104, 2007.

[69] F. Pozzi, T. Di Matteo, and T. Aste. Centrality and peripherality in filtered graphs from dynamical financial correlations. *Advances in Complex Systems*, 11:927–950, 2008.

[70] T. Di Matteo, F. Pozzi, and T. Aste. The use of dynamical networks to detect the hierarchical organization of financial market sectors. *Eur. Phys. J. B*, 73:3–11, 2010.

[71] D. Y. Kenett, M. Tumminello, A. Madi, G. Gur-Gershgoren, R. N. Mantegna, and E. Ben-Jacob. Dominating clasp of the financial sector revealed by partial correlation analysis of the stock market. *PLoS ONE*, 5:e15032. doi:10.1371/journal.pone.0015032, 2010.

[72] D. Y. Kenett, X. Huang, I. Vodenska, S. Havlin, and H. E. Stanley. Partial correlation analysis: applications for financial markets. *Quantitative Finance*, 15:569–578, 2008.

[73] T. Aste, W. Shaw, and T. Di Matteo. Correlation structure and dynamics in volatile markets. *New J. Phys.*, 12:085009, 2010.

[74] M. McDonald, O. Suleman, S. Williams, S. Howison, and Neil F. Johnson. Impact of unexpected events, shocking news, and rumors on foreign exchange market dynamics. *Phys. Rev. E*, 77:046110, 2008.

[75] G. Buccheri, S. Marmi, and R. N. Mantegna. Evolution of correlation structure of industrial indices of U.S. equity markets. *Phys. Rev. E*, 88:012806, 2013.

[76] J. P. Onnela, A. Chakraborti, K. Kaski, and J. Kertész. Dynamic asset trees and black monday. *Physica A*, 324:247–252, 2003.

[77] W. Jang, J. Lee, and W. Chang. Currency crises and the evolution of foreign exchange market: Evidence from minimum spanning tree. *Physica A*, 390:707, 2010.

[78] H. Kaya. Eccentricity in asset management. *Journal of Network Theory in Finance*, 1:45–76, 2015.

[79] V. Tola, F. Lillo, M. Gallegati, and R.N. Mantegna. Cluster analysis for portfolio optimization. *J. Econ. Dyn. Control*, 32:235–258, 2008.

[80] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész, and A. Kanto. Dynamics of market correlations: Taxonomy and portfolio analysis. *Phys. Rev. E*, 68:056110, 2003.

[81] F. Pozzi, T. Di Matteo, and T. Aste. Spread of risk across financial markets: better to invest in the peripheries. *Sci. Rep.*, 3:1665, 2013.

[82] J. Riordan. *An Introduction to Combinatorial Analysis*. New York, Wiley & Sons, 1958.

[83] W. M. Song, T. Di Matteo, and T. Aste. Nested hierarchies in planar graphs. *Discrete Appl. Math.*, 159:2135, 2011.

[84] S. Boccaletti, G. Bianconi, R. Criado, C. Del Genio, J. Gómez-Gardeñes, and et al. M. Romance. The structure and dynamics of multilayer networks. *Phys Rep*, 544:1–122, 2014.

[85] F. Battiston, V. Nicosia, and V. Latora. Metrics for the analysis of multiplex networks. *Phys. Rev. E*, 89:032804, 2013.

[86] V. Nicosia and V. Latora. Measuring and modelling correlations in multiplex networks. *Phys. Rev. E*, 92:032805, 2015.

[87] M. Szell, R. Lambiotte, and S. Thurner. Multirelational organization of large-scale social networks in an online world. *Proc. Natl. Acad. Sci. USA*, 107 (31):13636–13641, 2010.

[88] P. Klimek and S. Thurner. Triadic closure dynamics drives scaling laws in social multiplex networks. *New J. Phys.*, 15 (6):063008, 2013.

[89] B. Corominas-Murtra, B. Fuchs, and S. Thurner. Detection of the elite structure in a virtual multiplex social system by means of a generalised k-core. *PLos One*, 9:e112606, 2014.

[90] T.G. Kolda, B.W. Bader, and J.P. Kenny. Higher-order web link analysis using multilinear algebra. *ICDM 2005: Proc. of the 5th IEEE International Conference on Data Mining*, page 242–249, 2005.

[91] G.A. Barnett, H.W. Park, K. Jiang, C. Tang, and I.F. Aguillo. A multi-level network analysis of web-citations among the world's universities. *Scientometrics*, 99 (1):5–26, 2014.

[92] Z. Wu, W. Yin, J. Cao, G. Xu, and A. Cuzzocrea. Community detection in multi-relational social networks. *Web Information Systems Engineering WISE 2013, in: Lecture Notes in Computer Science, Springer Berlin Heidelberg*, 8181:43–56, 2014.

[93] S. Poledna, J.L. Molina-Borboa, M. van der Leij, S. Martinez-Jaramillo, and S. Thurner. Multi-layer network nature of systemic risk in financial networks and its implications. *Journal of Financial Stability*, 20:70–81, 2015.

[94] M. Montagna and C. Kok. Multi-layered interbank model for assessing systemic risk. *Kiel Working Paper No. 1873*, 2013.

[95] R. Burkholz, M. V. Leduc, A. Garas, and F. Schweitzer. Systemic risk in multiplex networks with asymmetric coupling and threshold feedback. *Physica D*, 2015.

[96] I. Rivals, L. Personnaz, L. Taing, and M.C. Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23 (4):401–407, 2007.

[97] M. Tumminello, S. Miccichè, F. Lillo, J. Varho, J. Piilo, and R. N. Mantegna. Community characterization of heterogeneous complex systems. *J. Stat. Mech.*, P01019, 2011.

[98] M. Tumminello, S. Miccichè, F. Lillo, J. Piilo, and R.N. Mantegna. Statistically validated networks in bipartite complex systems. *PLoS ONE*, 6:e17994.doi:10.1371/journal.pone.0017994, 2011.

[99] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2:193–218, 1985.

[100] Y. Liu, P. Cizeau, M. Meyer, C.K. Peng, and H.E. Stanley. Correlations in economic time series. *Physica A*, 245:437–440, 1997.

[101] T. Andersen and T. Bollerslev. Intraday periodicity and volatility persistence in financial markets. *Journal of Empirical Finance*, 4:115–158, 1997.

[102] K. Pearson. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, 58:240–242, 1895.

[103] F. Pozzi, T. Di Matteo, and T. Aste. Exponential smoothing weighted correlations. *Eur. Phys. J. B*, 85:6, 2012.

[104] S.R. Nanda, B. Mahanty, and M. K. Tiwari. Clustering indian stock market data for porfolio management. *Expert System with Applications*, 37:8793–8798, 2010.

[105] G. A. V. Pai and T. Michel. Clustering indian stock market data for porfolio management. *Evolutionary Optimization of Constrained K-means Clustered Assets for Diversification in Small Portfolios*, 13:1030–1053, 2009.

[106] M. Tumminello, F. Lillo, and R.N. Mantegna. Kullback-Leibler distance as a measure of the information filtered from multivariate data. *Phys Rev E Stat Nonlin Soft Matter Phys*, 76:031123, 2007.

[107] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30:81–93, 1938.

[108] C. Spearman. The proof and measurement of association between two things. *American Journal of Psychology*, 15:72–101, 1904.

[109] N. Musmeci, T. Aste, and T. Di Matteo. Relation between financial market structure and the real economy: Comparison between clustering methods. *PLoS ONE*, 10 (4):e0126998. doi: 10.1371/journal.pone.0126998, 2015.

[110] G. Meissner. *Correlation risk modeling and management*. Wiley, 2014.

[111] H. Markowitz. Portfolio selection. *The Journal of Finance*, 7 (1):77–91, 1952.

[112] A. J. McNeil, R. Frey, and Paul Embrechts. *Quantitative Risk Management. Concepts, Techniques and Tools*. Princeton Series in Finance, 2005.

[113] J. Daníelsson. *Financial risk forecasting*. Wiley, 2011.

[114] J. Hakala and U. Wystup. *Foreign Exchange Risk*. Risk Publications, 2002.

[115] A. Jacquier and S. Slaoui. Variance dispersion and correlation swaps. *Working Paper, Imperial College London*, 2010.

[116] G. Vidyamurthy. *Pairs Trading, Quantitative Methods and Analysis*. John Wiley & Sons, 2004.

[117] I. Nelken. Variance swap volatility dispersion. *Derivatives Use, Trading & Regulation*, 11 (4):334, 2006.

[118] B. F. King. Market and industry factors in stock price behavior. *Journal of Business*, 39:139–190, 1966.

[119] D.J. Fenn, M.A. Porter, S. Williams, M. McDonald, N.F. Johnson, and N.S. Jones. Temporal evolution of financial market correlations. *Phys. Rev. E*, 84:026109–026121, 2011.

[120] G.J. Ross. Dynamic multifactor clustering of financial networks. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 89:022809, 2014.

[121] T. Preis, D. Y. Kenett, H. E. Stanley, D. Helbing, and E. Ben-Jacob. Quantifying the behavior of stock correlations under market stress. *Scientific Reports*, 2:752, 2012.

[122] Eugene F Fama and Kenneth R French. The capital asset pricing model: theory and evidence. *Journal of Economic Perspectives*, pages 25–46, 2004.

[123] Roger Lowenstein. *Origins of the Crash: The Great Bubble and Its Undoing*. Penguin Books, 2004.

[124] M. Baily and D. Elliott. The us financial and economic crisis: where does it stand and where do we go from here? *The Initiative on Business and Public Policy at Brookings*, 2009.

[125] D. Kwiatkowski, P.C.B. Phillips, P. Schmidt, and Y. Shin. Testing the null hypothesis of stationarity against the alternative of a unit root. *Journal of Econometrics*, 54 (1–3):159–178, 1992.

[126] Louis Bachelier. Théorie de la spéculation. *Gauthier-Villars*, 1900.

[127] R. Cont. Empirical properties of asset returns: stylized facts and statistical issues. *Quantitative Finance*, 1:223–236, 2001.

[128] A. Clauset, C. R. Shalizi, and M. E. J. Newman. Power-law distributions in empirical data. *SIAM Rev.*, 51 (4):661–703, 2009.

[129] J. C. Hull. *Options, futures and derivatives*. Pearson Education, 1997.

[130] L. Kullmann, J. Töyli, J. Kertész, A. Kanto, and K. Kaski. Characteristic times in stock market indices. *Physica A*, 269:98–110, 1999.

[131] D. N. Joanes and C.A. Gill. Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society (Series D): The Statistician*, 47 (1):183–189, 1998.

[132] E. F. Fama. Efficient capital markets: a review of theory and empirical work. *The Journal of Finance*, 25:383–417, 1970.

[133] F. Black. Studies of stock price volatility changes. *Proceedings of the 1976 Meetings of the American Statistical Association*, page 171–181, 1976.

[134] A. Christie. The stochastic behavior of common stock variances: Value, leverage and interest rate effects. *Journal of Financial Economics*, 10:407–432, 1982.

[135] J. Yu. On leverage in a stochastic volatility model. *Journal of Econometrics*, 127:165–178, 2005.

[136] V. Plerou, P. Gopikrishnan, L. A. Nunes Amaral, X. Gabaix, and H. E. Stanley. Economic fluctuations and anomalous diffusion. *Phys. Rev. E*, 62:R3023(R), 2000.

[137] F. Galton. Regression towards mediocrity in hereditary stature. *Journal of the Anthropological Institute of Great Britain and Ireland*, 58:246–263, 1886.

[138] J. F. Kenney and E. S. Keeping. *Mathematics of Statistics*. Princeton, NJ: Van Nostrand, 1947.

[139] J. M. Wooldridge. *Introductory Econometrics*. South Western, 2013.

[140] M. Tumminello, C. Coronnello, F. Lillo, S. Miccichè, and R. N. Mantegna. Spanning trees and bootstrap reliability estimation in correlation-based networks. *Int. J. Bifurcat. Chaos*, 17:2319–2329, 2007.

[141] C.E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilitá. *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 1936.

[142] O.J. Dunn. Estimation of the medians for dependent variables. *Annals of Mathematical Statistics*, 30 (1):192–197, 1959.

[143] O.J. Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56 (293):52–64, 1961.

[144] A. Edelman. Eigenvalues and condition numbers of random matrices. *SIAM J. Matrix Anal. Appl.*, 9, 1988.

[145] A. M. Sengupta and P. P. Mitra. Distributions of singular values for some random matrices. *Phys. Rev. E*, 60:3389, 1999.

[146] E. J. Elton and M. J. Gruber. *Modern Portfolio Theory and Investment Analysis*. J. Wiley and Sons, 1995.

[147] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization. *Eur. Phys. J. B*, 27:DOI: 10.1140/epjb/e20020153, 277–280, 2002.

[148] Z. Burda and J. Jurkiewicz. Signal and noise in financial correlation matrices. *Phys. Rev. E*, 344 (1-2):67–72, 2004.

[149] T. Guhr and B. Kälbe. A new method to estimate the noise in financial correlation matrices. *Journal of Physics A: Mathematical and General*, 36 (12), 2004.

[150] G. Livan, S. Alfarano, and E. Scalas. Fine structure of spectral properties for random correlation matrices: An application to financial markets. *Phys. Rev. E*, 84:016113, 2011.

[151] R. Litterman and K. Winkelmann. Estimating covariance matrices. *Goldman Sachs, Risk Management Series*, 1998.

[152] J. Papenbrock and P. Schwendner. Handling risk-on/risk-off dynamics with correlation regimes and correlation networks. *Financ Mark Portf Manag*, 29:125–147, 2015.

[153] D. Y. Kenett, T. Preis, G. Gur-Gershgoren, and E. Ben-Jacob. Quantifying meta-correlations in financial markets. *EPL*, 99:38001, 2012.

[154] C. Coronnello, M. Tumminello, F. Lillo, S. Miccichè, and R. N. Mantegna. Sector identification in a set of stock return time series traded at the London Stock Exchange. *Acta Physica Polonica*, 36:2653–2680, 2011.

[155] S. Cavaglia, J. Diermeirer, V. Moroz, and S. Zordo. Investing in global equities. *Journal of Portfolio Management*, 30:88–94, 2004.

[156] S. Heston and K. Rouwenhorst. Does industrial structure explain the benefits of international diversification? *Journal of Financial Economics*, 36:3–27, 1994.

[157] M. A. Ferreira and P. M. Gama. Correlation dynamics of global industry portfolios. *J. of Multi. Fin. Manag.*, 20:35–47, 2010.

[158] G. Bekaert, R. Hodrick, and X. Zhang. International stock return comovements. *Journal of Finance*, 64:2591–2626, 2009.

[159] T. W. Epps. Comovements in stock prices in the very short run. *Journal of the American Statistical Association*, 74:291–298, 1979.

[160] M. Munnix, R. Schafer, and T. Guhr. Impact of the tick-size on financial returns and correlations. *Physica A*, 389:4828–4843, 2010.

[161] Y. Shapira, Y. Berman, and E. B.-Jacob. Modelling the short term herding behaviour of stock markets. *New Journal of Physics*, 16:053040, 2014.

[162] B. Tóth and J. Kertész. The epps effect revisited. *Quantitative Finance*, 9 (7):793–802, 2009.

[163] K. Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 11:559–572, 1901.

[164] T. Aste, R. Gramatica, and T. Di Matteo. Exploring complex networks via topological embedding on surfaces. *Phys. Rev. E*, 86:036109, 2012.

[165] O. Borůvka. O jistém problému minimálním (about a certain minimal problem). *Práce mor. přírodověd. spol. v Brně III*, 3:37–58, 1926.

[166] M. Bastian, S. Heymann, and M. Jacomy. Gephi : An open source software for exploring and manipulating networks. *AAAI Publications, Third International AAAI Conference on Weblogs and Social Media*, 2009.

[167] R.L. Graham and P. Hell. On the history of the minimum spanning tree problem. *Annals of the History of Computing*, 7 (1):43–57, 1985.

[168] N. Vandewalle, F. Brisbois, and X. Tordoir. Non-random topology of stock markets. *Quantitative Finance*, 1:372–374, 2001.

[169] R. Pastor-Satorras, A. Vázquez, and A. Vespignani. Epidemic spreading in scale-free networks. *Phys. Rev. Lett.*, 86:3200, 2001.

[170] J. B. Kruskal. On the shortest spanning subtree of a graph and the traveling salesman problem. *Proceedings of the American Mathematical Society*, 7:48–50, 1956.

[171] R.C. Prim. Shortest connection networks and some generalizations. *Bell System Technical Journal*, 36 (6):1389–1401, 1957.

[172] M.L. Fredman and R.E. Tarjan. Fibonacci heaps and their uses in improved network optimization algorithms. *Journal of the ACM*, 34 (3):596, 1987.

[173] B. Chazelle. A minimum spanning tree algorithm with inverse-Ackermann type complexity. *Journal of the Association for Computing Machinery*, 47 (6):1028–1047, 2000.

[174] A. Garas, P. Argyrakis, and S. Havlin. The structural role of weak and strong links in a financial market network. *Eur. Phys. J. B*, 63:265–271, 2008.

[175] C. K. Tse, J. Liu, and F. C.M. Lau. A network perspective of the stock market. *Journal of Empirical Finance*, 17:659–667, 2010.

[176] H.-J. Kim, I.-M. Kim, Y. Lee, and B. Kahng. Scale-free network in stock markets. *Journal of the Korean Physical Society*, 40 (6):1105–1108, 2002.

[177] J.-P. Onnella, K. Kaski, and J. Kertész. Clustering and information in correlation based financial networks. *Eur. Phys. J. B*, 38:353–362, 2004.

[178] N. J. Foti, J. M. Hughes, and D. N. Rockmore. Nonparametric sparsification of complex multiscale networks. *PLoS ONE*, 6 (2):e16431, 2011.

[179] T. Aste. An algorithm to compute Planar Maximally Filtered Graphs (PMFG), 2012.

[180] K. Kuratowski. Sur le probléme des courbes gauches en topologie. *Fund. Math.*, 15:271–283, 1930.

[181] G.P.Massara, T. Aste, and Di Matteo T. Network filtering for big data: Triangulated maximally filtered graph. *Journal of Complex Networks (in press)*, 2016.

[182] G. Bonanno, F. Lillo, and R.N. Mantegna. High-frequency cross-correlation in a set of stocks. *Quantitative Finance*, 1 (1):96–104, 2001.

[183] K. Baba, R. Shibata, and M. Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46:657–664, 2004.

[184] L. Kaufman and P. J. Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L_1-Norm and Related Methods*, pages 405–416, 1987.

[185] J. MacQueen. Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 1:281–297, 1967.

[186] R. Rammal, G. Toulouse, and M. A. Virasoro. Ultrametricity for physicists. *Rev. Mod. Phys.*, 58:765, 1986.

[187] S. Fortunato. Community detection in graphs. *Physics Reports*, 486:75–174, 2010.

[188] B. Everitt. *The Cambridge Dictionary of Statistics*. Cambridge University Press, 1998.

[189] S. Wagner and D. Wagner. Comparing clusterings - an overview. *Technical Report, ITI Wagner, Faculty of Informatics, Universität Karlsruhe (TH)*, 2007.

[190] B. Efron. Bootstrap methods: Another look at the jackknife. *Ann. Stat.*, 7:1–26, 1979.

[191] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and their Application*. Cambridge University Press, 1997.

[192] A Guide to Industry Classification Benchmark. http://www.icbenchmark.com/.

[193] N. Musmeci, T. Aste, and T. Di Matteo. Risk diversification: a study of persistence with a filtered correlation-network approach. *Journal of Network Theory in Finance*, 1:1–22, 2015.

[194] D. Chetalova, R. Schäfer, and Thomas Guhr. Zooming into market states. *J. Stat. Mech.*, page P01029, 2014.

[195] D. J. Fenn, M. A. Porter, P.J. Mucha, M. McDonald, S. Williams, N.F. Johnson, and N.S. Jones. Dynamical clustering of exchange rates. *Quantitative Finance*, 12:1493, 2012.

[196] T.G. Andersen, T. Bollerslev, F.X. Diebold, and P. Labys. Forecasting realized volatility. *Econometrica*, 71 (2):579–625, 2003.

[197] M. Kritzman, Y. Li, S. Page, and R. Rigobon. Principal components as a measure of systemic risk. *Journal of Portfolio Management*, 37:112–126, 2011.

[198] Z.Y. Zheng, B. Podobnik, L. Feng, and B.W. Li. Changes in cross-correlations as an indicator for systemic risk. *Sci. Rep. 2*, page 888; DOI:10.1038 srep00888, 2011.

[199] H. Kunsch. The jackknife and the bootstrap for general stationary observations. *Ann. Statist.*, 17:1217–1241, 1989.

[200] K. A. Spackman. Signal detection theory: Valuable tools for evaluating inductive learning. *Proceedings of the Sixth International Workshop on Machine Learning. San Mateo, CA: Morgan Kaufmann*, page 160–163, 1989.

[201] S. Pafka and I. Kondor. Noisy covariance matrices and portfolio optimization II. *Physica A*, 319:487 – 494, 2003.

[202] R.A. Fisher. On the 'probable error' of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

[203] D. N. Politis and H. White. Automatic block-length selection for the dependent bootstrap. *Econometrics Reviews*, 23 (1):53–70, 2004.

[204] J. M. Hilbe. *Logistic Regression Models*. Chapman & Hall/CRC Press, 2009.

[205] R. Schmidt and U. Stadmüller. Nonparametric estimation of tail dependence. *Scandinavian Journal of Statistics*, 33:307–335, 2006.

[206] P. Fiedor. Networks in financial markets based on the mutual information rate. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, 89:052801, 2014.

[207] T. You, P. Fiedor, and A. Hołda. Network analysis of the shanghai stock exchange based on partial mutual information. *Journal of Risk and Financial Management*, 8:266–284, 2015.

[208] S. V. Buldyrev, R. Parshani, G. Paul, H. E. Stanley, and S. Havlin. Catastrophic cascade of failures in interdependent networks. *Nature*, 464(7291):1025–1028, 2010.

[209] E. Goffman. *Frame Analysis: An Essay on the Organization of Experience*. Harvard University Press, 1974.

[210] B. Corominas-Murtra and S. Thurner. The weak core and the structure of elites in social multiplex networks. In Antonios Garas, editor, *Interconnected Networks*, pages 165–177. Springer International Publishing, 2016.

[211] L. Tang, X. Wang, and H. Liu. Community detection via heterogeneous interaction analysis. *Data Min. Knowl. Discov.*, 25 (1):1–33, 2012.

[212] Catastrophic cascade of failures in interdependent networks. *Nature*, 464 (7291):1025–1028, 2010.

[213] C.D. Brummitt, R.M. D'Souza, and E.A. Leicht. Suppressing cascades of load in interdependent networks. *Proc. Natl. Acad. Sci. USA*, 109 (12):E680–689, 2012.

[214] A. Cardillo, M. Zanin, J. Gómez-Garde nes, M. Romance, A. García del Amo, and S. Boccaletti. Modeling the multi-layer nature of the european air transport network: Resilience and passengers re-scheduling under random failures. *Eur. Phys. J. Spec. Top.*, 215 (1):23–33, 2013.

[215] M. Barigozzi, G. Fagiolo, and G. Mangioni. Identifying the community structure of the international-trade multi-network. *Physica A*, 390 (11):2051–2066, 2011.

[216] L. Bargigli, G. Di Iasio, L. Infante, F. Lillo, and F. Pierobon. The multiplex structure of interbank networks. *Quantitative Finance*, 15 (4):673–691, 2015.

[217] G. S. Shieh. A weighted kendall's tau statistic. *Statistics & Probability Letters*, 39:17–24, 1998.

[218] F. Lindskog, A. McNeil, and U. Schmock. Kendall's tau for elliptical distributions. In Georg Bol, Gholamreza Nakhaeizadeh, Svetlozar T. Rachev, Thomas Ridder, and Karl-Heinz Vollmer, editors, *Credit Risk*, chapter 8, pages 149–156. Springer, 2003.

[219] M. R. C. van Oordt and C. Zhou. The simple econometrics of tail dependence. *Economics Letters*, 116 (3):371–373, 2012.

[220] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41, 3:1–58, 2009.

[221] P. Holme and H. Saramäki. Temporal networks. *Physics Reports*, 519:97–125, 2012.

[222] M. Tumminello, F. Lillo, and R. N. Mantegna. Hierarchically nested factor model from multivariate data. *EPL*, 78:30006, 2007.

[223] C. E. Shannon. A mathematical theory of communication. *Bell Syst. Tech. J.*, 27:379, 1948.

[224] R. Diestel. *Graph Theory ed. 3*. Springer-Verlag, 2005.

[225] P. Hall, J. L. Horowitz, and B.-Y. Jing. On blocking rules for the bootstrap with dependent data. *Biometrika*, 82 (3):561–574, 2004.